Webシラバス統合のためのレコード解析

伊東, 栄典
九州大学情報基盤センター

山田, 信太郎
九州大学大学院システム情報科学府

廣川, 佐千男
九州大学情報基盤センター

出版情報：SIG-SWO. 第1回 (A201-05), pp.1-7, 2002-11. 人工知能学会セマンティックウェブとオントロジー研究会
バージョン：
権利関係：

人工知能学会研究会資料
SIG-SWO-A201-05

Webシラバス統合のためのレコード解析

伊東栄典† 山田信太郎* 廣川佐千男†
{itou@cc, yamashin@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

Web で公開されたシラバスデータの収集と統合のための研究を行なっている。シラバスには科目名，担当者，講義目標，教科書などの項目が共通に含まれる。しかしその表現は，大学，学部，学科ごとに異なっている。これらを統合するには，個別のシラバス・データをレコードと捉え，どのフィールドがどの項目に対応するかを識別する必要がある。シラバスデータの半構造性に着目し，まず，大学，学部，学科ごとに共通テンプレートを抽出し，次にテンプレートのインスタンスにおける各フィールド内容の出現頻度に基づき，フィールドの属性推定を行なう方式を提案する。

An analysis of records in syllabus web pages

Eisuke Itoh† Shintaro Yamada* Sachio Hirokawa†
{itou@cc, yamashin@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

A lot of syllabus pages are available as web pages in many educational organizations. The authors consider these syllabus as resources for Web Mining, and try to realize an knowledge base system about syllabus for educational application. The syllabus has an abstract information of each lecture. By integrating syllabus of a specific field, they can be a knowledge resource for selecting good textbooks and for comparison of feature of educational organizations.

The authors had proposed an approach to extract records and fields. That approach is based on frequent appearing structure over syllabus pages, such as link structure, or HTML tags patterns.

1 はじめに

Web の普及により，多種多様な情報が Web 上に公開されるようになっており，そのページ数は膨大な数になっている。この膨大な数のページから利用者に有用な情報を捜し出すための検索システムや，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 [12] 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 [11, 14] 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, 2]。それから更に，意味的に同種の不均質な情報を統合・組織化して用いる為の研究が進んでいる。

本研究では，具体的な対象テーマとして，Web 上に存在するシラバス情報を統合して扱う事を目的としている。そのために，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行っている [7, 15, 16, 17]。このシステムの実現により，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分野 (科目) に関する情報の調査の支援が期待できる。

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。Web 上に公開されているシラバスを統合するために，以下の五つの手順で研究を進めている。

(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

2 シラバスページとシラバス一覧

大学などの高等教育機関が提供するシラバスは，科目名の一覧を記述した目次となるページ (シラバス一覧) と，個々の科目の内容を詳細を記述したページ (シラバス・ページ) から構成される。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，シラバス情報を提供するサイトは，図 1 に示すリンク構造をもっていることが多い [9, 16, 17]。

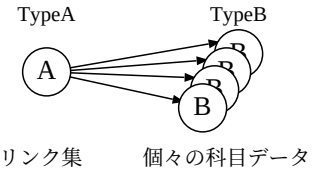

図 1: リンク構造

このとき，科目内容を記述した B のシラバスページは，同一サイト内では同じ記述形式 (書式) が用いられる場合が殆んどである。しかし，複数のシラバスサイトを比較すると，それが同じ大学内のサイトであったとしても，学部や学科でが変わると記述形式が変わってしまう場合が多い。どのサイトでもシラバスとしての記述内容は類似しているが，記述の方法が統一されていない。

シラバス情報を統一的に表現する為に必要な項目名 (field-name) を，約 50 のサイトのシラバスから調べた [15]。同じ属性を持つが属性名が異なる場合のために，一つの属性名を代表的に扱う事にしている。例えば,「授業科目名」を表すための属性名には「授業科目名，授業科目，テーマ，研究主題，講義科目，科目名」がある。この属性を表す場合,「授業科目名」を属性名として代表させることにした。

他の属性も同様に調べ，共通計画表と名付けたシラバスを扱うためのメタデータ形式を作成した。表 1 に作成した共通計画表の内容を示す。

表 1: 共通計画表

代表属性名	対応する属性名
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

3 レコード抽出

個々の科目の内容を記述したシラバスページは，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」，レコード内の箱で囲まれた単位を「フィールド」と呼ぶ。

「フィールド」は更に,「属性名 (attribute-name)」,「属性値 (attribute-value)」,「その他」の三つに分けられる。この分け方は，フィールドを扱う目的に依り，どこに当たるかの割り当ては変化する。本稿ではシラバス情報を統合す

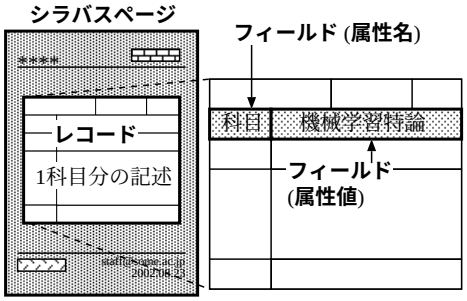

図 2: ページ・レコード・フィールド

ることを目的としているため，’担当教官’, ’授業科目名’ といった文字列は属性名として扱われる。それらの属性名に対応した文字列が属性値として扱われる。

複数のシラバスサイトの統合に役立つ情報を抽出するためは，集めたシラバス・ページから，「属性名」と，それに対応する「属性値」を高精度で抽出し，一つの科目についての情報であるレコードを構成する必要がある。

レコードを構成する場合，一つのページ内に複数のレコードが記述されている場合と，一つのページ内に一つのレコードしか記述されていない場合を考慮する必要がある。収集した具体的な例を見ると，一つのシラバス・ページには，一つのレコードしか記述されていない場合が大半であった。そこで，本稿では「ページ = レコード」として，分析を行っている。

3.1 テンプレート抽出

シラバス・ページからレコードおよびフィールドを抽出するために，シラバスを記述している枠組 (テンプレート) を抽出する方法を用いている [7]。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枠組で記述されていることが多い。そこで,「同一サイト内にあるシラバス・ページは，ほぼ同じ枠組を持つ」との前提に基づき，同一サイト内にある複数のシラバス・ページが共通して持つ枠組 (HTML の構造) を，テンプレートとして抽出する。

HTML 構造の表現としては，木構造を用いるのが一般的であるが，二つの木の構造的同等性を手間がかかる [2, 5, 10]。シラバスの場合，シラバス一覧 (A 型) のページからリンクされているページは，下位になるシラバス・ページ (B 型) か，学科のトップページなどの上位となるページしかない。しかも，これらのうち B 型のページは，同一のワープロ文書から生成されたものや，データベース等から自動的に生成されたものが多く，従って HTML の構造が同一になっている場合が多い。

そこで，我々は HTML 木の比較ではなく，タグ列として比較することにより枠組構造を提案している [8, 13]。同一サイト内に存在する B 型のシラバス・ページ群に，共通して出現するタグ列 (タグパターン) を捜し出す。最もマッチしたタグ列 (タグパターン) が，テンプレート構造となる。この方法を用いることで，効率良く B 型のシラバス・ページ群から，枠組構造となるテンプレートを抽出する事が出来る。実際に実験してみた所，大半のシラバスページ群については，この簡易な方法でテンプレートが抽出可能であり [7]，大まかなシラバス情報を集めるという事を目的とする場合には，この方法で十分であると考えている。

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，HTML 的に考えると入れ子構造 (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。

4 属性名・属性値の推定

前節に記述した方法で，HTML で記述されたシラバス・ページから個々のフィールドを切り出す部分まで，ある程度の精度で実現できている。しかしながら，各フィールドが属性名なの

か属性値なのか，それ以外なのかを判別する必要がある。また，属性名と，それに対応する属性値を結びつける必要がある。これらを実現するために，以下に述べる二つの方法を行った。

4.1 方法 1:特徴的キーワードを用いた方法

まず，単純な推定方法として，シラバスに出現する特徴的な単語 (キーワード) を用いて，{ 属性名/属性値/その他 } を判定する方法を考えた。キーワードとしては，表 1 の共通計画表に含まれる単語を用いた。その方法は以下の通りである。

- 共通計画表に含まれる単語がフィールドに出現した場合，そのフィールドは「属性名」であると判定する。
- その直後に出てくるフィールドが共通計画表に含まれない単語である場合，前に出た「属性名」に対応する「属性値」と判定する。

例えばフィールドに出現する文字列が,「科目名　物理学　担当　伊東　…」である場合を考える。‘科目名’ という文字列は，共通計画表の最初のカテゴリに含まれるため，この部分は「属性名」であろると判定する。次の ‘物理学’ は表に含まれない文字列であるため,「属性値」であると判定する。

しかしながら，二つの問題がある。一つは，シラバス・ページの記述の仕方が問題になる。次節では，図 4 を例に用いているが，このページの様に,

学年	授業科目名	科目コード	…
4 年	応用数学 1	３１４０１	…

と表の中で属性名が並び，その後属性値が連続する場合には，{ 属性名/属性値/その他 } を判定する事が出来ない。

また，いくつかのページでは，‘科目名’ という記述なしに ‘物理学’ などの具体的な属性値が記述されている場合がある。この場合には属性名が出現しないため，そのフィールドの属性を特定できない。また ‘概要’ のように，比較的長い文章の中には ‘細目’ のような事なる属性名が表れる事もある。

上記の理由により，方法 1 ではフィールドから「属性名」と「属性値」を推定することが十分に出来ない。

4.2 方法 2：属性名と属性内容の識別

一つのフィールドの属性を，一つのレコード内にあるフィールドの文字列だけで推定することについては限界がある。そこで，本節では，同一形式のレコードにおいて，ひとつのフィールドにどのようなインスタンスがあるかという，大局的な分析を用いてフィールドの属性判定を行う方法を提案する。

例として，阿南工業高等専門学校における電気工学科平成 13 年度のシラバス一覧 (図 3) について説明する。[1]

図 3: シラバス一覧

このページは 44 個のリンクを含む。そのうち 3 個は，電気工学科のホームページ，各学科のシラバス一覧のページ，ならびに高専のトップページへのリンクである。残りの 41 個のリンク

[1] http://www.anan-nct.ac.jp/gakka/syllabus/h13/curri_e.html

は,「応用数学 1」,「応用数学 2」,「電子工学」…などの具体的な科目シラバスへのリンクである。これら 41 個の HTML ファイルは共通のタグ列により表現され，タグにより囲まれる空でないテキスト部分は 31 個ありレコードのフィールド数と考えられる (図 4)。

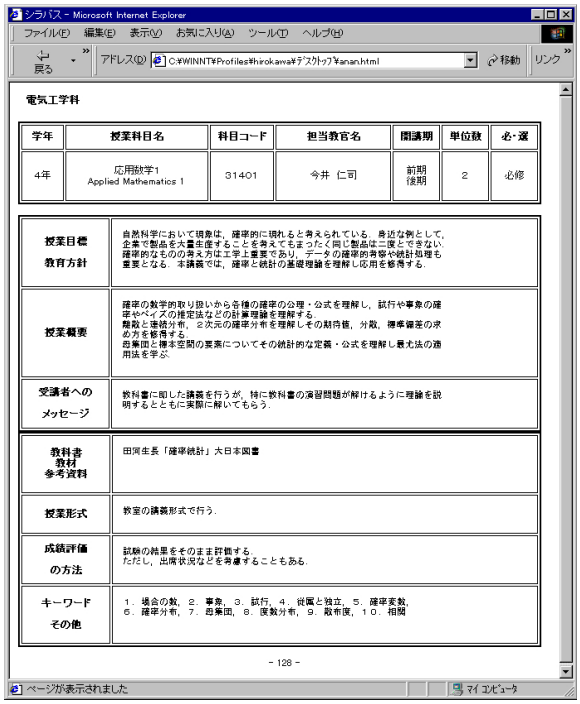

図 4: シラバス・ページ中のフィールド

表 2 は，各フィールドの内容が 41 個のシラバス・ページ中にどれだけ共通して現れていたかを表す。例えば，3 番目のフィールドは，41 個のシラバス全て共通で「学年」が現れていたことを表す。このように出現数が 41 であるフィールドは，共通計画表の属性名に対応しているか，あるいは，学科名などの共通的なものである。表 2 においては，出現頻度が 10 回以上のものしか書いていない。一方，11,12,13 番目のフィールドに授業科目名，科目コード，担当教官名など，それぞれのシラバスに応じて異なる個別の情報があられているので，出現頻度は非常に低い。これを見ると，フィールド値は

表 2: フィールド出現頻度

出現数	番号	内容
41	1	シラバス
41	2	電気工学科
41	3	学年
41	4	授業科目名
41	5	科目コード
41	6	担当教官名
41	7	開講期
41	8	単位数
41	9	必・選
10	10	５年
15	10	４年
28	14	前期 後期
14	15	１
24	15	２
17	16	選択
24	16	必修
41	17	授業目標 教育方針
41	19	授業概要
41	21	受講者への メッセージ
41	23	教科書 教材 参考資料
41	25	授業形式
41	27	成績評価 の方法
41	29	キーワード その他

N: ほとんどすべてのレコードにおいて同じフィールド値になっているもの

V: レコードごとにフィールド値がばらばらで出現頻度が低いもの

の二種類に分けられる。N の部分が属性名，V の部分が属性値と見なすことができる。さらに V の部分は出現頻度が低いだけでなく，文字列の長さも N の部分と比べて大きいことが分かる。

図 4 のシラバス・ページの上部分では,「学年，授業科目，科目コード，担当教官名，開講期，単位数，必・選択」のように属性名が連続して横に並び，その後に同じ数だけ属性値が並んでいる。$\Sigma = \{N, V\}$ をアルファベットとする文字列として表すと，この部分は，$NNNNNNNVVVVVVV$ となる。一般的には N^nV^n のように N が n 個ならび直後に V が n 個ならぶパターンになっている。この N と V の並びにおいて，i 番目の属性名 N に対する属性値は，i 番目の V に現れている。「授業目標，教

育方針」,「授業概要」あるいは「受講者へのメッセージ」のように長いテキストからなる部分では，$n = 1$ となり，縦に並ぶ。

図 5 は，各フィールドに現れる文字列の出現頻度と長さをプロットしたものである。x 軸フィールド番号，左の y 軸が出現頻度，右の y 軸が文字列長である。左半分に着目すると，1 番から 9 番までのフィールは短い文字列からなり，全てのレコードにおいて同じものとなっていることが分かる。10 番から 16 番までのフィールドは文字列長がある程度あり，出現頻度はほとんどが 1,2 回程度で，41 回というものはない。ただし，3〜9 番のフィールド群と 10〜16 番フィールド群が対応している。学年，学期，単位の様に 20 科目以上で属性値が共通になるものもある。右半分については，頻度が 41 回の部分と，頻度が 1 回の部分が交互に現れている。高頻度のフィールドの文字列長は短く，低頻度のフィールドの文字列長は極端に長いことが分かる。

このように，フィールド・インスタンスの出現頻度と文字列長により，属性名と属性値の識別ができる。さらに，一連のパターンにおける位置の対応により，属性名と属性値の対応も推定できる。

5 おわりに

Web 上に公開されているシラバスページ群から，一つの科目についての情報を，レコードとして抽出するための方法について検討した。既に開発しているフィールドの切り分け方法を用いて，各フィールドの属性名・属性値の判定および属性名と属性値の対応付けを行う方法について検討した。二つの方法を提案し，それぞれについて検討した。

方法 1 として，共通計画表で拾い挙げた属性名の文字列をマッチさせる簡易な方法を用いた。しかしながら，この方法では，属性名が横に並ぶ表の場合には，うまく推定ができない。そこで，方法 2 として，シラバス・ページ群の多くに出現するフィールドを「属性名」と推定し，さらにフィールドの順序関係を考慮して横に並ぶ表の場合にも「属性名」と「属性値」の推定を行う方法を考えた。しかしながら，方法 2 を用いても，属性名のフィールド無しに属性値が出現する場合はうまく推定できていない。

今後は，シラバス情報を統合的に利用できるシステムの充実を進める予定である。また，そのシステムを利用して，利用者に有益な知識を提供するための方法を検討し，実装して提供していきたいと考えている。レコード抽出や,「属性名」と「属性値」の判定および対応付けについても，多くのデータで方法の有効性を実証していきたい。

また，本稿で提案した方法は，図 1 の構造を持つページ群を自動的に収集・統合することができると考えられる。シラバス以外にも，新聞記事やグルメ情報，観光情報などのような，情報も統合利用するシステムも構築して行きたい。

参考文献

[1] C. C. Aggarwal, F. Al-Garawi and P. S. Yu : "*Intelligent Crawling on the World Wide Web with Arbitrary Predicates*", Proc. WWW2001.

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

[3] 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.

[4] 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)

[5] 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.

[6] 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.

[7] 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.

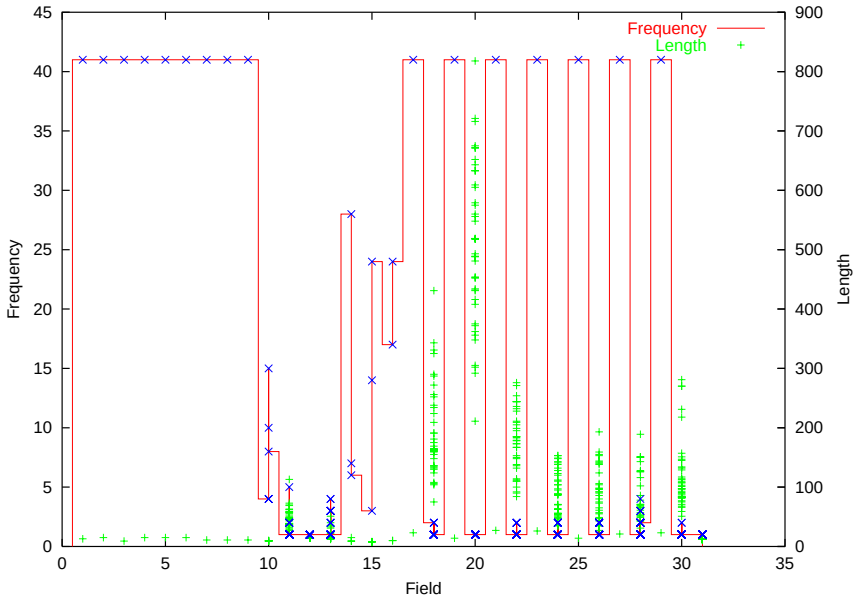

図 5: 出現頻度による属性名・属性値の識別

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 山田信太郎, 伊東栄典, 廣川佐千男 : “自動収集した Web シラバスデータの分析と考察,” 情報科学技術フォーラム 2002 (FIT2002), 一般講演論文集第 4 分冊 (N-32), pp.301-302, Sep.,2002.