Auto DB construction from Web syllabus

伊東, 栄典
九州大学情報基盤センター

松永, 吉広
九州大学システム情報科学研究院

山田, 信太郎
(株)NEC ソフトウェア九州

廣川, 佐千男
九州大学情報基盤センター

出版情報：人工知能学会全国大会論文集. 17, pp.48-49, 2003-06. 人工知能学会
バージョン：

1D4-08

Webシラバスからの DB 構成

Auto DB construction from Web syllabus

伊東 栄典 *1 Eisuke Itoh
松永 吉広 *2 Yoshihiro Matsunaga
山田 信太郎 *3 Shintaro Yamada
廣川 佐千男 *1 Sachio Hirokawa

*1 九州大学情報基盤センター Computing and Communications Center, Kyushu Univ.
*2 九州大学システム情報科学研究院 Dept. of Comp. Sci. and Com. Eng., Kyushu Univ.
*3 (株)NEC ソフトウェア九州 NEC Software Kyusyu Inc.

1. はじめに

近年，情報技術の発達と，ネットワーク環境の普及，情報機器の低価格化とともに，さまざまな分野で情報技術の利用が進んでいる．教育の分野も例外ではなく，多くの高等教育機関で，教材やシラバスデータの電子化が進み，主に Web を介して利用されるようになってきている [2]．電子化され Web 上で公開される情報は，教材などの実際の教育に使われるコンテンツだけではなく，シラバスや受講状況などの授業についての情報も電子化されつつある．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ため，系統的な利用は困難である．

また一方，Web の広がりに伴い，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 [3, 4] 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の収集分類についての研究 [5] 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ても，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].

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・ページ群を収集し，科目概要，教科書等の検索が可能なシラバス DB を開発を目指している [6, 7, 8]．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．その実現のためには，効率的なシラバスページ収集，ページからのレコード部分抽出，DB への格納，具体的な知識提示といった機能を実現する必要がある．本稿では，知識を格納蓄積するシラバス DB の構築に関して考察する．

2. シラバス DB 構築

2.1 メタデータ形式

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．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けれども，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．

そのため，同一種類の内容を示すのに，各組織ごとに異なる属性名を用いる場合が多い．例えば，講義を担当する講師を表すために,「担当教官，担当，担当者，教官名，担当教員」といった属性名が用いられている．このため，ただシラバス・ページを収集しただけでは，系統的な利用は困難である．

表 1: 共通計画表

共通属性名	対応属性名
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

そこで，シラバス情報を統一的に表現する為に必要な属性名を，約 50 のサイトのシラバスから調べた [6]．その結果，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た．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.

2.2 レコード抽出

各科目の内容を記述した Web 上のシラバスページから，その内容を DB に格納するためには，ページの HTML テキストから，具体的な内容を表す部分を抽出しなければならない．図 1 に，ページからレコードを抽出する場合の切り分け方を示す.

レコード部分の抽出には，多くのページに繰返し出現する構造を利用する [7, 8]．多くのシラバスサイトは，図 2 に示す構造をしている．科目一覧となるページ (A 型) があり，そこから個々の科目内容を示すページ (B 型) へリンクが繋がった構造である．個々の科目についての内容を記述した B 型ページは，概ね同じ HTML 構造を持つ．これを利用し，同一サイトの B 型ページ群に頻出する HTML タグ構造を抜き出すことで，シラバスの内容を HTML テキストから抽出する方法を考案した.

A: 伊東栄典, 九州大学情報基盤センター, 〒 812-8581 福岡市東区箱崎 6-10-1, Tel.092-642-4037, Fax.092-642-3844, itou@cc.kyushu-u.ac.jp

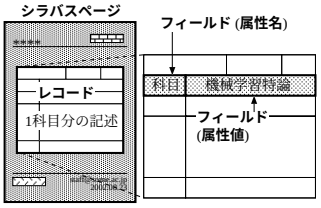

図 1: レコード・フィールド

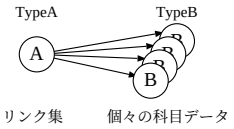

図 2: シラバスサイトのリンク構造

2.3 シラバス XML スキーマ

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．関係 DB にデータを格納する場合，スキーマが必要である．表 1 に示した共通計画表の形で，DB に格納する事は可能である．

現在，大学評価・学位授与機構 (NIAD) ではシラバスデータを記述するための XML スキーマを提供している [9, 10]．我々の決めた共通計画表には 10 項目しか属性がない．これに対し，NIAD 提供のシラバス XML スキーマは，30 以上の属性名を持ち，より詳細な形での記述が可能であるように定義されている．さらに，繰返し記述も可能であるし，XML であるため今後の拡張も可能である．そこで，共通計画表の形で格納するのではなく，シラバス XML スキーマに変換できる形で DB に格納する事にした．

表 2 に，シラバス DB で用いる DB スキーマを示す．共通計画表の属性名の対応も併記した．殆んどの属性名は，NIAD のシラバス XML スキーマにある属性名である．「キーワード」という属性名だけは，共通計画表に有りるが，シラバス XML スキーマに無い．そこでシラバス DB スキーマに `keyword` という属性名を追加した．

表 2: シラバス DB スキーマ

シラバス XML スキーマの属性名	「共通計画表」の対応する属性名
SyllabusType	
Common	
Courses	
CommonType	
institution	
faculty	
department	
program	
academicYear	
CourseType	
code	授業コード
title	授業科目名
eTitle	授業科目名
year	
termSystem	
term	授業学期
day	曜日と時間
time	曜日と時間
requiredSelective	
credit	単位数
classType	
room	
Lectures	
abstract	概要
keyword	キーワード
Plan	概要
prerequisiteCompetences	
prerequisiteCources	関連科目
corequisiteCources	
courseObjectives	
evaluation	評価方法
textbooks	教材
references	教材
remarks	
LectureType	
name	担当教官
:	
PlanType	
Session	
preparation	
topics	概要
assignment	

3. 試作システム

図 3 に，現在構築中のシステムの概要を示す．システムは，Web ページ収集部，シラバス判定部，レコード抽出部，DB 格納部の 4 つからなる．現在，それぞれの部分は，独立して動作するようになっている．今後はシステムの統合を進める予定である．

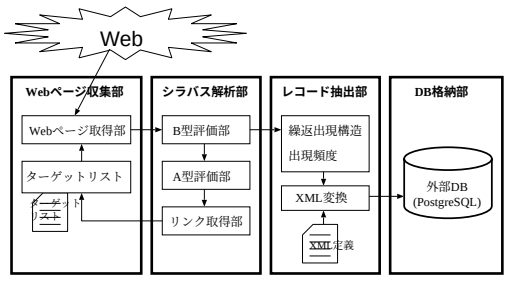

図 3: システム概要図

4. おわりに

本稿では，Web 上に公開されているシラバス・ページからのシラバス DB の構成について述べた．今後は Web から自動的に情報を収集する統合したシステムとして，実装を進めていく予定である．

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] 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)

[3] 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.

[4] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom.: "*The TSIMMIS Project: Integration of Heterogeneous Information Sources*". Proc. of IPSJ Conf., pp.7-18, 1994.

[5] 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.

[6] 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.

[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 SIG-KBS-A202, pp.59-64, Sep., 2002.

[8] 伊東栄典, 山田信太郎, 廣川佐千男 : "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, pp.(05-1)–(05-7), 2002.

[9] 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), pp2003.

[10] `http://svrrd2.niad.ac.jp/syllabus/10/syllabus10.xsd`