3P-4

大学シラバスデータベースの設計と試作

島松千春† 伊東栄典‡ 廣川佐千男‡ 篠原正典*

† 九州大学理学部物理学科情報理学コース,
‡ 九州大学情報基盤センター , * メディア教育開発センター

1 はじめに

情報技術の発達と，ネットワーク環境の普及により様々な分野で情報技術の利用が進んでいる．教育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 [10]．メディア教育開発センター (NIME) では，電子教材に関するポータルサイト [4] を作成し，電子化教材の公開と普及を行っている．また，教育情報ナショナルセンター (NICER) では [7]，様々な教育情報を収集・公開している．

大学評価や大学改革では教育内容についての情報が重要になってきている．シラバスは大学の教育内容を表す重要な情報であると同時に，学術分野全体を表す知識ベースであると言える．

著者らは Web マイニングについての研究を行なってきており，その技術を利用した Web シラバスの収集・統合による教育情報データベース構築を目指している [2, 3, 5, 8, 9]．本発表では，構築中のシラバスデータベースについて述べる．

2 関連研究

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，系統的な利用は困難である．そこで，各組織が独自に公開している Web 上のシラバス文書群を収集し，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．

井田，野澤らは，大学評価や教育内容の分析に用いるために，シラバスの統合とそこからの知識発見について研究している [12, 6]．また，彼らはシラバスを記述するための XML スキーマを開発し，大学評価・学位授与機構 (NIAD) から公開している [6]．

青野らは，内容が類似している半構造化データ群の統合についての研究を行なっており，その例としてシラバスの統合を検討している [11]．

3 大学シラバスデータベース

本研究では，大学が公開するシラバスを収集・統合し，知識ベースとしての利用を目指している．その実現のためには，シラバスの効率的な発見・収集，Web ページ群からのレコード部分の抽出，DB への統合，具体的な知識提示手法の開発が必要である．それぞれの機能について，以下で説明する．

3.1 発見・収集

Web からのシラバス発見および収集 [9] について述べる．まず，Google 等の一般検索サイトを用いて数十サイトから Web シラバス群を収集した．収集したページを分析し，シラバスの特徴を現す単語（特徴語）を抽出した．次に，抽出した特徴語を用い，与えたページがシラバスであるかどうかを判定する，判定関数を作成した．

次に，クローリングにより Web シラバスの収集を行った．教育機関の Web サイトをクロールし，集めたページがシラバスであるかどうかを前述の判定関数により判定した．

クローリングの収集の開始 URL は，文部科学省のサイトにある，国立大学，公立大学，国公私立短期大学，私立大学，国立高等専門学校のサイトへのリンク集ページを用いた．これらには国内高等教育機関 1,320 校へのリンクが存在している．

なお，収集対象としているファイルは，文書解析プログラムの都合から HTML と PDF 文書だけに限っている．Word 文書や一太郎文書といった形式の文書は対象としていないため，これらの文書ファイルの収集は行っていない．

上記のシステムを用い，現在までに 282,344 個のシ

ラバス文書ファイルを発見・収集している．なお，そのうちの HTML 文書は 144,026 個である．

3.2 抽出・統合

次に Web シラバス文書群からのレコード抽出と統合 [2, 5] について述べる．

著者らは，シリーズ型の HTML 文書群から，レコード部分を抽出する手法について研究開発している [1]．「シリーズ型」とは，特定の様式に基づいて作成された，同一サイト内に存在するページ群のことを指す．Web シラバスは組織毎に様式が決まっており，その様式に基づく文書ファイルが科目数分存在するという，典型的なシリーズ型の文書群である．そのため，開発した手法を用いることで，レコード部分となるテキストを抽出することができる．

また，抽出したシラバスを，NIAD シラバス XML スキーマへ統合することの研究も試みている [5]．様々な様式で書かれたシラバスを，一つの特定の様式に統合することで，検索や統計といった知識抽出のための処理が容易になる．

3.3 検索

検索については，具体的なシステムが未だ出来ていない．そのため，ここでは検索システムの構想だけを述べる．まず最初に，利用者が入力した検索語を含む科目を表示するといった，従来の検索システムと同様のシステムを開発する．

次に，統計的な処理をおこなう検索システムを考える．調査・分析をしている研究者からの具体的な要求としては，国際関係について教育している組織の数を調べたい，電子教材を公開している組織の数を調べたい，といった事がある．これらにの要求を実現する検索システムが必要である．

他にも，知識発見を行う検索システムも開発する．我々は，「Matrix 検索」と名づけた多面的分析システムを開発している．また，「概念グラフ」と名づけた，文書群からの知識発見システムを研究開発している．これらを用いることで，大量の文書群からの知識発見が可能になると考えている．

4 おわりに

本稿では，Web 上に存在するシラバスを収集し，データベースとして統合することについて述べた．また，それを利用した検索システムについて考察も行った．今後は，本稿で提案したシステムの実装を進める予定である．

参考文献

[1] Hirokawa, S., Itoh, E. and Miyahara, T.: Semi-Automatic Construction of Metadata from A Series of Web Documents, *LNAI 2903, Proc. of AI2003*, pp. 942–953 (2003).

[2] Kuboyama, T., Miyahara, T., Hirokawa, S. and Itoh, E.: Information Extraction from Web Pages Using Semi-Structured Data Alignment, *Proc. 9th World Multi-Conference on Systemics, Cybernetics and Informatics* (2005).

[3] Matsunaga, Y., Yamada, S., Ito, E. and Hirokawa, S.: A Web Syllabus Crawler and Its Efficiency Evaluation, *Proc. of International Symposium on Information Science and Electrical Engineering 2003*, pp. 565–568 (2003).

[4] メディア教育開発センター：教育メディアポータルサイト．http://www.ps.nime.ac.jp/.

[5] 伊東栄典，竇ギョク峰，廣川佐千男：情報処理学会マルチメディア，分散，協調とモバイル (DICOMO 2004) シンポジウム論文集，pp. 345–348 (2004).

[6] 井田正明，野澤孝之，芳鐘冬樹，宮崎和光，　喜多一：シラバスデータベースシステムの構築と専門教育課程の比較分析への応用，大学評価・学位研究，No. 2, pp. 87–97 (2005).

[7] 教育情報ナショナルセンター (NICER)：http://www.nicer.go.jp/.

[8] 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).

[9] 篠原正典，地蔵真作：Web 上の高等教育に役立つコンテンツの自動収集・抽出－授業シラバスの自動抽出－，JSiSE　第 30 周年記念全国大会講演論文集，pp. 247–248 (2005).

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

[11] 平野健太郎，青野雅樹：情報系科目を用いた HTML シラバスの XML 変換と内容分析，電子情報通信学会 SIG Notes WI2-2005-28〜49，pp. 83–88 (2005).

[12] 野澤孝之，井田正明，芳鐘冬樹，宮崎和光，　喜多一：シラバスの文書クラスタリングに基づくカリキュラム分析システムの構築，情報処理学会論文誌，Vol. 46, No. 1, pp. 289–300 (2005).