Ｗｅｂシラバス統合による教育情報ライブラリ構築

○伊東栄典*1, 島松千春*2, 廣川佐千男*1, 篠原正典*3
*1 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, E-mail:itou@cc.kyushu-u.ac.jp
*2 九州大学理学部物理学科情報理学コース
*3 メディア教育開発センター

概 要

高等教育機関でも情報化が進んでいる．大学でもシラバスを Web 上に公開することが行なわれている．教育機関のシラバス群は，その機関が提供する教育全体も表している．更には，日本全国のシラバスば，現代日本の大学における教育情報ライブラリであるといえる．

著者らは Web 上のシラバスを収集・統合した教育情報ライブラリの構築を目指している．本稿では，著者らが進めている Web シラバス統合システムについて述べる．自己開発のトピッククローラーにより，388 ドメイン (大学) から約 18 万のシラバス文書ファイルを収集した．集めたシラバス文書のうち，HTML で記述されたファイルから具体的なシラバスのテキストを抽出・統合するシステムを試作した．また簡単なキーワード検索を行なうシステムも試作した．

更に集めたシラバス数について，統計的な分析を行なった．収集ファイル数は，ほぼ Zipf の法則に従っていることが分かった．ファイル数上位の大学は，大学全体でシラバス DB を構築しており，全ての学部学科の授業情報をその DB で提供していることがわかった．また，ファイル数が 100 個程度の場合は学部単位であり，20 個程度の場合は学科単位であることも分かった．

キーワード：Web, シラバス, データマイニング, データ統合, 知識ベース

Web syllabi integration for construction of educational knowledge library

Eisuke Itoh*1, Chiharu Shimamatsu*2, Sachio Hirokawa*1, Masanori Shinohara*3
*1 Computing and Communications Center, Kyushu University.
Hakozaki 6-10-1, Higashi-ku, Fukuoka, 812-8581, Japan.
Phone: +81-92-623-4037, Fax: +81-92-642-3844, E-Mail: itou@cc.kyushu-u.ac.jp
*2 Faculty of Science, Kyushu University.
*3 National Institute of Multimedia Education

Abstract

Information and communication technologies change university education. A lot of syllabus pages are available as web pages in many educational organizations. The authors consider these syllabus as resources for Web Mining. The syllabi are not only abstract information of contents and also academic knowlege base.

The authors are constructing educational knowledge library by web syllabi integration. They developed a prototype of web syllabi integration system and also constructed a simple keyword search engine for integrated syllabi data.

They applied frequency analysis for number of colleced web-syllabi, and found that number of web-syllabi follows zipf's law. They also studied correlation analysis between number of web-syllabi and the number of lectures, and students.

keywords : Web, syllabi, data mining, data integration, knowledge base

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が進んでいる．様々な人がインターネットを利用し，Web 上には膨大な数の Web ページが存在している．そこから我々は求める情報を世界中から集めることができる．求めている情報が記述されたページを探す手段としては，Google(`http://www.google.com/`) などの検索エンジンがある．

Web ページを文書とみなし，多数の文書から知識を得る Web マイニングについても様々な研究されている．また，ページ群をデータベースのように用いる，データ統合技術についても研究されている．Web 上には，データを扱う際のルールや記述様式が定まっていない Web ページが大量に存在している．そういったページをデータベースのように統合できれば，多様な知識ベースを構築できる．

教育分野においても，情報技術の利用が進んでいる．e ラーニングなどの電子教材や，Web シラバスなど，多くの高等教育機関で Web を介した情報提供が行なわれるようになってきている [11]. 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 [5] を作成し，電子化教材の公開と普及を行なっている．また，教育情報ナショナルセンター (NICER) [8] では，様々な教育情報を収集・公開している．

各大学の自発的な教育情報の公開も進んでいる．米国マサチューセッツ工科大学 (MIT) が始めた OCW(Open Course Ware) が始めた教育内容の公開は，日本にも波及しており，日本 OCW 連絡会 (`http://www.jocw.jp/`) では現在 7 つの大学が教育内容の公開を行なっている．JOCW 加盟の大学数は徐々に増加すると思われる．

OCW では教材を電子的に公開することもある．様々な大学が，多数の電子教材を公開すれば，それを集めることで大学の電子教材を蓄積した教育情報ライブラリになるだろう．

現在の図書館では，具体的な教育内容を蓄積しているわけではない．今後，電子教材が普及すれば，教材の蓄積は進むと考えられ，その教材にはメタデータが付与されることになるだろう．教育の具体的な内容をコンテンツとすれば，コンテンツに対するメタデータの一部がシラバスとなるだろう．

また，一つの教育機関のシラバス群は，その機関が提供する教育全体も表している．大学評価や大学改革では，大学が提供する教育内容についての情報が重要である．シラバスは，大学の教育内容を表す資料であり，大学評価や単位認定の場合には重要な根拠資料となる．また，各大学の教育内容の特徴比較にも用いることができる．更には，日本全国のシラバスを収集できれば，その中には現代日本の大学で行なわれている学術知識全体が保持されることになるだろう．

本稿では，Web 上に公開されている大学のシラバスを統合した，教育情報ライブラリ構築について述べる．次節で，統合に際しての問題点と，関連研究について述べる．

2 Web 統合システム概要

本研究では，大学が公開するシラバスを収集・統合し，教育情報ライブラリとして利用することを目指している．その実現のためには，シラバスの効率的な発見・収集，シラバスファイル群からのレコード部分の抽出および DB への統合，具体的な知識検索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．

2.1 発見・収集

知識ベース構築のためには，シラバスを大規模に収集する必要がある．シラバス収集の確実な方法として，各大学にシラバスデータを提供してもらうという手段がある．しかしこの手法は膨大な時間と労力がかかってしまう．一方，クローラー技術を用いて Web 上で公開されているシラバスを収集するという手段もある．我々は，Web シラバスを効率的に収集するクローラーを開発し [9, 10]，Web シラバスファイルの収集を行なっている．Web からのシラバス発見および収集について述べる．

まず最初に，Web シラバスの特徴分析を行なった．Google 等の一般検索サイトを用いて数十サイトから Web シラバス群を収集する．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．クローリングの開始 URL は，文部科学省のリンク集ページ[1]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 1,230 校へのリンクが存在している．

なお，現在の所 HTML 文書と PDF 文書だけを収集対象にしている．MS-Word や一太郎といった形式の文書は対象としていない．これは，シラバスの判定で用いている文書解析プログラムの都合からである．Word や一太郎の内部データ形式を解釈し，内部の文字列を扱うことができるならば，シラバス判定関数は適用可能である．

2.2 抽出・統合

収集した Web シラバス文書群から統一のとれた DB を構築するためには，レコード抽出と統合が必要である．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 [3, 4].「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．

Web 上で公開されていシラバスは，各大学，学部，学科など各組織がそれぞれ個別に作成したものである．シラバスは，大まかな記述様式は存在しているものの，詳細な部分については統一されていない．そのため，ただ集めてキーワード検索を行なうだけでは，系統的な利用は困難である．系統的に利用するためには，各組織が独自に作成した Web 上のシラバス文書群を収集し，科目名，科目概要，教科書などの項目を指定検索が可能であるように統合する必要がある．

そこで，抽出したシラバスを NIAD シラバス XML スキーマ [12] へ統合することの研究も試みている [6]．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．

[1] http://www.mext.go.jp/b_menu/link/main_b12.htm

2.3 検索

検索については，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簡単な検索システムは試作している．しかしながら詳細な分析を行なうための検索システムは実現していない．ここでは検索システムの構想を述べる．

簡単な検索システムに続いて，統計的な処理をおこなう検索システムを考える．教育に関する調査・分析をしている研究者からは，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，といったの具体的な要求を聞いている．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．

3 関連研究

Web からのシラバス収集については，トピッククローラーの技術を用いている．トピッククローラーは，特定のトピックに関する文書のみを収集するソフトウェアである．トピッククローラーについては，Aggrawal らの研究が先駆的である [1]．Chakrabarti らも，Focused Crawler と名付けたトピッククローラーについて研究を行なっている [2]．

シリーズ型文書群からのレコード抽出については，梅原，岩沼らが行なっている [13, 14]．梅原，岩沼らは，シリーズ型文書における構造の類似性を利用して，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している．シリーズ型文書の特徴の一つに，記述様式 (テンプレート) の類似がある．同一組織のシラバス文書は，一つの記述様式にそって作成されている．そのため，シラバス文書群からのレコード抽出については，シリーズ型文書の特徴である記述様式の一意性を利用して行なっている．

シラバスの統合については，いくつかの研究が行なわれている．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 [17, 7]．また，シラバスを記述するための XML スキーマを開発し [7]，大学評価・学位授与機構 (NIAD) から公開している．また，青野らも，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っており，その例としてシラバスの統合を検討している [15, 16]．

4 システムの試作

試作したシステムの全体像を図 1 に示す．

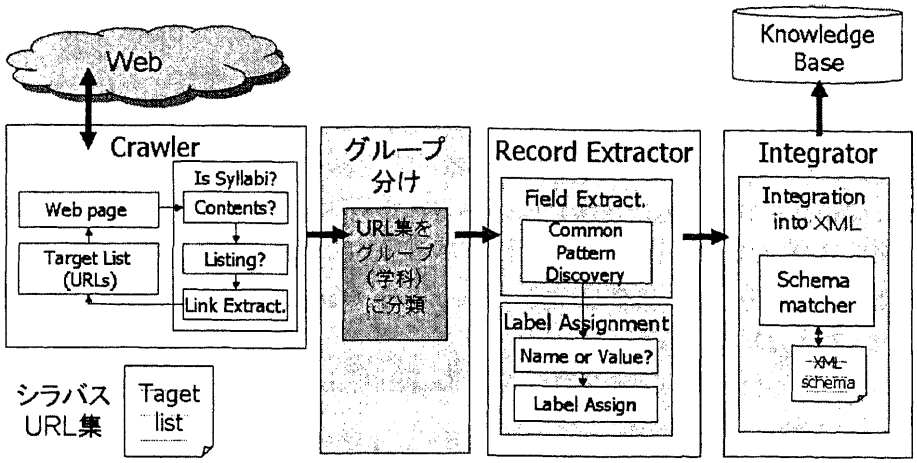


図 1: システム全体像

4.1 発見・収集

まず，2.1 節で述べたトピッククローラーを用い，現在までに 179,496 個のシラバス文書ファイルを発見・収集している．そのうちの HTML 文書は 159,196 個で PDF 文書は 20,300 個である．表 1 にファイル数，ホスト数およびドメイン数を示す．

表 1: 収集シラバスファイル数

収集ファイル総数	179,496
PDF ファイル数	20,300
HTML ファイル総数	159,196
(静的な HTML ページ)	93,391
(動的な HTML(CGI) ページ)	65,778
ホスト数	1,430
ドメイン数	388

4.2 グループ分類

続いて，集めたシラバスファイル群を，グループ毎に分類する．分類方法としては，組織単位，科目単位，専門単位，年度単位など，様々な方法がある．組織にも，大学・学部・学科・専攻といった階層がある．まずは，最も小さい組織単位をグループとして分類することを考えている．

最初の手法として，URL の文字列だけを見て，グループに分類する方法を行なった．その方法を述べる．静的な HTML ページの URL は，`http://HOST/PATH/FILE.htm` といった形をしている．

HOST と PATH の部分が同一で，FILE の部分のみ異なる URL は，同じグループのファイルだと考えられる．

CGI プログラム等により生成される動的な HTML ページは，`http://HOST/PATH/cgi_file?k1=v1&k2=v2…` といった形をしている．'?' 以降の文字列は，キー k1 に対する値が v1, キー k2 に対する値が v2, という意味を表しており，出現順序は関係しない．そこで，キーの文字列で並べ，その後，`http://HOST/PATH/cgi_file/v1/v2` のように値のみを並べた文字列に変更した．こうすることで，静的な HTML ページの URL 群を分類した方法で同様に分類できる．

上記の簡単な手法での分類した結果は，「同一テンプレートを持つ集団」をグループと定義しての分類結果となった．分類した結果を表 2 に示す．

表 2: 分類グループ数

	静的 HTML	動的 HTML(CGI)
グループ総数	2952	13191
平均 URL 数	33	9115
1URL のグループ	856	5

表 3: Top5 グループ (静的 HTML)

グループ名 (URL)	URL 数
www.tsc.u-tokai.ac.jp/risyuu_syllabus/	11855
kbweb3.kj.yamagata-u.ac.jp/2003/html/	2896
yoran-syllabus.kanagawa-u.ac.jp/2004/syllabus/	2361
yoran-syllabus.kanagawa-u.ac.jp/syllabus/	1854
www.ipc.kit.ac.jp/%7Estudent/syllabus/	1291

表 4: Top5 グループ (動的 HTML)

グループ名 (URL)	URL 数
wmt.bunri-u.ac.jp/syllabus/sylla_ichiran.php	9946
db.jm.hirosaki-u.ac.jp/cybouz/db.exe	5561
syllabus.sic.shibaura-it.ac.jp/syllabus/namazu/namazu.cgi	3927
sas.shonan.bunkyo.ac.jp/syll2004/show-kamoku.php	3341
eyume02.med.yamaguchi-u.ac.jp/syllabus_script/indexes/indexone.asp	2864

4.3 レコード抽出

グループに分類した Web シラバス文書群は，記述様式 (テンプレート) を同じくする文書群になっている．HTML で記述された Web シラバス文書群から，記述様式を抽出するプログラムを用い，その記述様式を使うことで内部のデータ (レコード) を抽出した．

抽出したレコードは，グループ単位で一つの CSV ファイルにまとめている．図 2 は，抽出により生成した CSV ファイルをエクセルで表示した様子を示している．

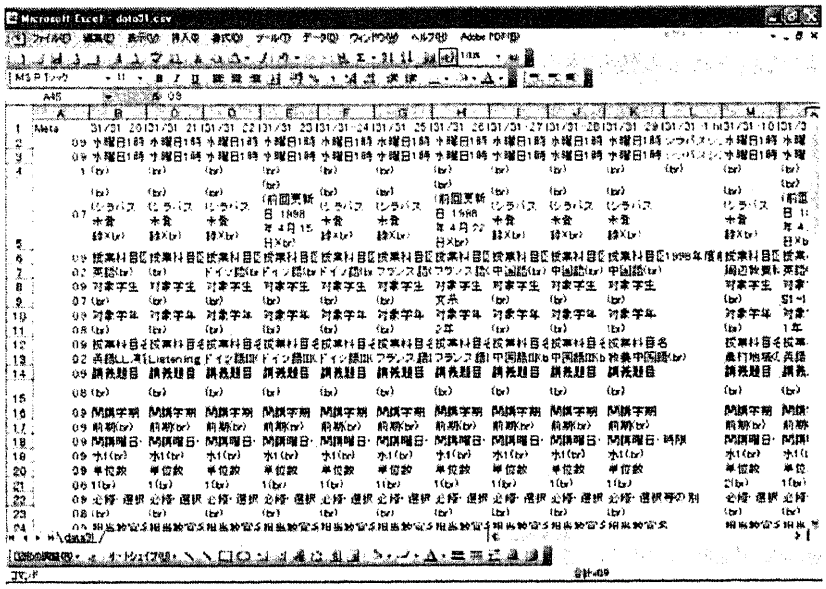


図 2: 抽出レコード

4.4 試作検索システム

シラバス文書群から抽出されたレコードは，グループ単位で一つの CSV ファイルにまとめてられている．このファイルを用いて，簡単なキーワード検索システムを試作している．図 3 に，検索の様子を示す．この例では「国際」の文字を含むシラバスを検索している．

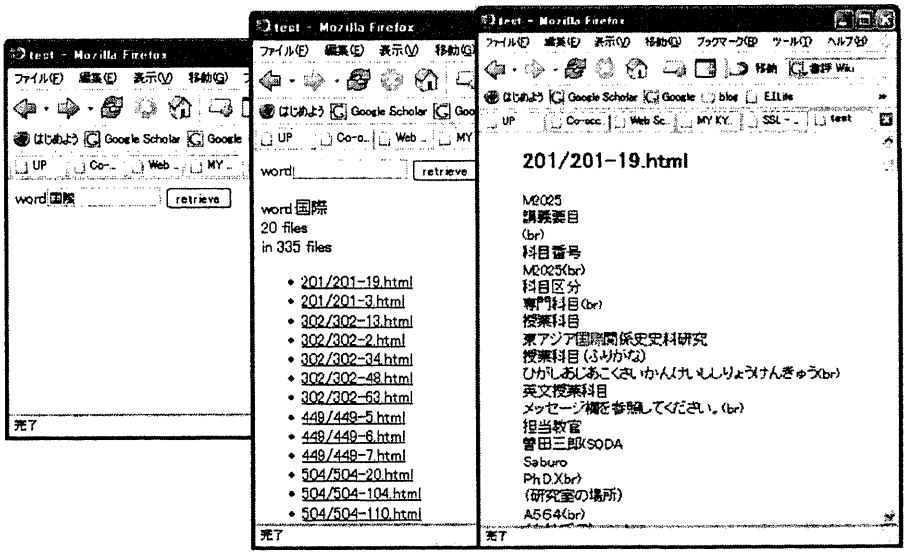


図 3: 検索システム

5 統計分析

収集したシラバスのファイル数が妥当なものであるかどうか検証するために，いくつかの統計分析を行なった．まず，ファイル数を，各グループ毎に，また各ドメイン (各大学) 毎にプロットした．ファイル数で降順に並べた順位を x 軸にとり，その順位のグループ/ドメインに含まれるファイル数を y 軸にプロットする．なお，どちらの軸も対数尺度にしている．

5.1 頻度分析

図 4 にグループ毎のファイル数を，図 5 にドメイン毎のファイル数をプロットした図を示す．どちらの図も，ほぼ Zipf の法則に従っていることを示している．

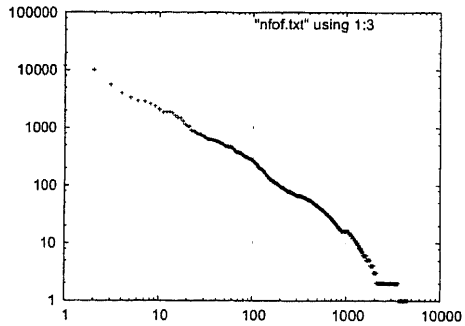


図 4: グループ毎のシラバスファイル数

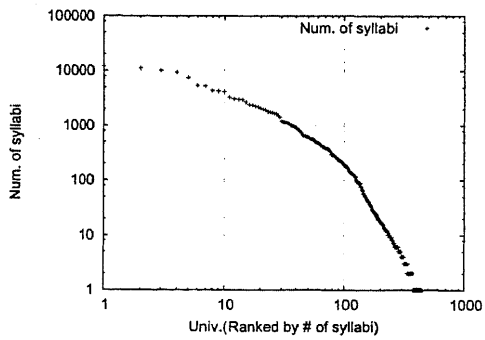


図 5: ドメイン毎のシラバスファイル数

図 4 について傾向を調べた．上位は 4 つのグループを詳細に調べた．続いて，ファイル数が 1000 個程度，500 個程度，100 個程度，20 個程度のグループをそれぞれ 5 つづつランダムに選び，それぞれについて詳細に調べた．また，ファイル数が 20 個未満の下位グループについても 10 グループを選び調査した．

上位 4 つのグループについては，一位の大学 (東海大学) は過去 6 年分のシラバスが含まれていることがわかった．その他の 3 つは，その他は大規模な大学の一年分のシラバスであった．学部や学科などに分けることなく，大学で提供されているシラバスは，全て一つのサイトから提供されていることになる．

1000 個程度のグループについても，すべて大学の一年分のシラバスであった．この場合も，学部や学科などに分けることなく，大学で提供されているシラバスは，全て一つのサイトから提供されていた．

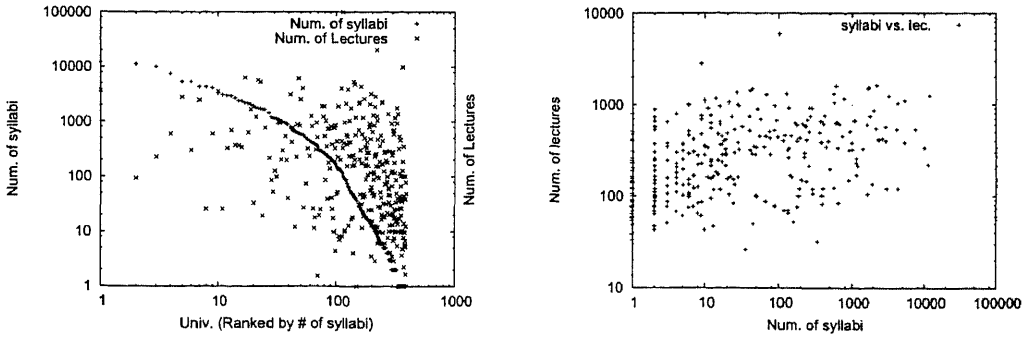
500 個程度のグループについては，学部学科など，一大学とは行かないまでも比較的大きな組織の一年分のシラバスであった．100 個程度のグループについては，学部や学科，専攻など，500 個程度のグループよりは小さな組織の 1 年分のシラバスであった．20 個程度のグループについては，学科や専攻など，100 個程度のグループよりもさらに小さな組織の 1 年分のシラバスであった．

なお，1 グループあたり 5URL 以下のグループについては，担当教官ごとに分かれているものが多かった．10 個前後のグループについては，特徴が分からなかった．20 個以上のグループは，組織の規模の違いはあれ，ほぼ確実に組織ごとに分けられていた．

5.2 相関解析

図 4 および図 5 は，どちらもほぼ Zipf の法則に従っているように思われる．シラバスファイル数と相関を，他のいくつかのデータと調べて見た．比較対象としては，大学の教員数 (教授・助教授・講師) および学生数を選んだ．この場合，グループ単位の教員数/学生は分からなかったため，ドメイン単位のシラバスファイル数との比較を行なった．

図 6 にシラバスファイル数と教員数をプロットした図を示す．(a) の方は図 5 に教員数の点もプロットしたものである．(b) の方では，一つの点は一つのドメイン (大学) に対応している．(x,y) 座標の値を，(シラバスファイル数, 教員数) として点をプロットしている．図 7 も同様に，学生数を用いてプロットをしたものである．



(a) (b)

図 6: シラバスファイル数と教員数

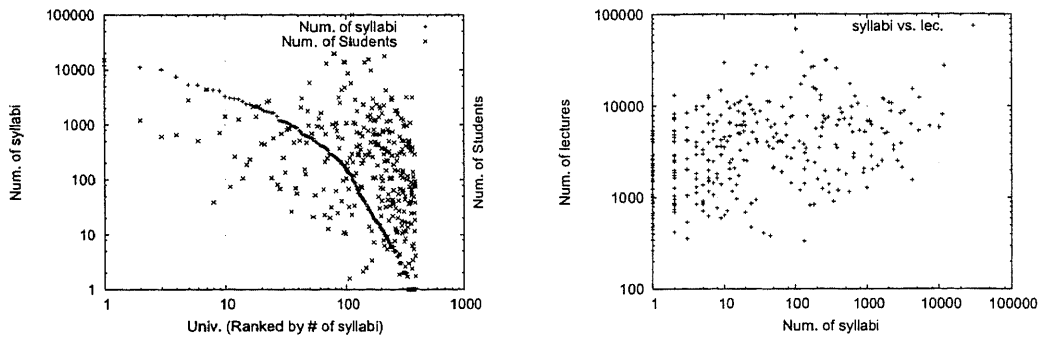


図 7: シラバスファイル数と学生数

図 6 と図 7 のどちらも，顕著な相関関係を示していない．何らかの特徴を見出すには，より詳細な分析が必要であろう．

6 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．実際に開発したクローラーを用いて，Web からシラバスを収集した．その結果，388 ドメイン (大学) から約 18 万のシラバス文書ファイルを収集できた．集めたシラバス文書のうち，HTML で記述されたファイルから具体的なシラバスのテキストを抽出・統合するシステムを試作した．また簡単なキーワード検索を行なうシステムも試作した．

集めたシラバスのファイル数について，統計的な分析を行なった．収集ファイル数は，ほぼ Zipf の法則に従っていることが分かった．ファイル数上位の大学は，大学全体でシラバス DB を構築しており，全ての学部学科の授業情報をその DB で提供していることがわかった．また，ファイル数が 100 個程度の場合は学部単位であり，20 個程度の場合は学科単位であることも分かった．

今後は，まず Web シラバス統合システムを完成させる予定である．また，検索方法について検討を行なう予定である．今回の試作システムでは簡単なキーワード検索しか行なっていない．項目名を指定しての検索，例えば科目名に含まれる文字列での検索など，を実現する予定である．また，作成したシステムを公開する予定である．

参考文献

[1] Aggarwal, C. C., Al-Garawi, F. and Yu, P. S.: Intelligent Crawling on the World Wide Web with Arbitrary Predicates, *Proc. WWW2001* (2001). http://www10.org/cdrom/papers/pdf/p110.pdf.

[2] Chakrabarti, S., Punera, K. and Subramanyam, M.: Accelerated Focused Crawling through Online Relevance Feedback, *Proc. WWW2002* (2002). http://www2002.org/CDROM/refereed/336/index.html.

[3] 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).

[4] 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).

[5] 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/.

[6] 伊東栄典，寶ギョク峰，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).

[7] 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).

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

[9] 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).

[10] 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).

[11] 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).

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

[13] 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，人工知能学会論文誌，　Vol. 16, No. 5, pp. 408–416 (2001).

[14] 梅原雅之，岩沼宏治，永井宏和：事例に基づくシリーズ型 HTML 文書の意味論理構造の自動認識，人工知能学会論文誌，　Vol. 17, No. 6, pp. 690–698 (2002).

[15] 平野健太郎，青野雅樹：DTD マッチングによる大学シラバスの相互変換，第 67 回情報処理学会全国大会 (2005).

[16] 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).

[17] 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).