Web 上に公開されたシラバス情報の自動収集

山田信太郎* 伊東栄典† 廣川佐千男‡
* 九州大学大学院システム情報科学府
† 九州大学情報基盤センター

Automatic collection of the Syllabus information exhibited on Web

Shintaro Yamada* Eisuke Itoh† Sachio Hirokawa†
*Graduate School of Information Science and Electrical Engineering, Kyushu University
†Computing and Communications Center, Kyushu University.

1. はじめに

情報技術の発達および社会の情報通信基盤の普及に伴い、教育の電子化が進んでいる [1]。国内の各大学では教材やシラバスデータも電子化され、情報ネットワーク、特に Web を介して利用されるようになってきている。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。

我々が自分の知らない新しい分野やテーマについて調べる場合、教科書の選定や学習する内容、講義順序を決めるために様々な情報を利用する。本研究では Web 上に公開されているシラバスから情報の抽出・統合を行い、その情報を利用して目的の分野に関する情報を整理し、提供する知識獲得システムの構築を目指している。

現在のシラバスデータは、各組織が自由に作成しているため、書式が統一されていない。このため、系統的な利用方法は困難である。一方、HTML データを代表とする半構造化データから知識を抽出する研究が進んでいる [2, 3, 4]。これらの技法を Web 上に公開されているシラバスからの知識獲得に用いる。手順は以下のようになる。

1. シラバスデータの性質分析
2. Web 上に公開されているシラバスデータの収集
3. HTML のシラバスデータをレコード項目への切り分け
4. 最後に切り分けられたデータを用いて知識を獲得

HTML のシラバスデータをレコード項目へと切り分けるためにはラッパー自動生成等の技術を用いる。知識の獲得では、教材名、著者、その講義のキーワード等の情報を取り出す。本論文では、各組織が Web 上に公開しているシラバスデータを自動的に収集する手法について考察する。自動収集の一次段階として、Web サーチエンジンを使ったキーワード検索と再帰的なデータ収集を用いる。この一次段階で収集したデータには多くのノイズが含まれている。このため、組織内のページデータのリンク構造 (サイトマップ) を用いてシラバスデータ部分を推測する。次に、シラバスに関連する単語データの出現頻度を用いて各ページの特徴抽出を行ない、シラバスデータの推定を行なう。実際に国内の教育機関 (*.ac.jp) 組織を対象に、データの収集と特徴抽出によるシラバスデータの推測を行なった結果についても報告する。

2. シラバスデータの分析

シラバスデータを収集するための準備段階として、現在のシラバスデータがどのような性質を持っているかを調べる。

2.1 シラバスの性質

収集対象であるシラバスデータの性質を調べる。具体的には、実際に公開されている 52 サイトのシラバスデータを人手により調査した。その結果、シラバスデータは多くの場合表の形式を取っていることが判明した [5]。ここでいう表とは、項目名と項目値のペアで構成されているものである。今後この表のことを授業計画表と呼ぶ。授業計画表は様々な項目名とそれに対応する項目値からなる。作り手が異なるシラバスでは、同じ項目値であっても項目名に異なる言葉が使われている場合が多い。例えば、「授業を担当する教官の名前」を項目値として持つ項目名には、「担当教官」や「担当者」、「教員名」などの様々な言葉が使われている。このように同じ項目値に複数の項目名が用いられているため、単純に項目名だけで情報を区別することはできない。このことが知識獲得のために項目値を利用する事を困難にしている。また、多くのシラバスに共通して現れる項目値というものが存在する。例えば「授業の名前」や「使用する参考書」などを表す項目値は、多くのシラバスに存在している。この、多数のシラバスに共通して現れる項目値を使うことで、作り手の異なるシラバスでも同じ授業計画表で表すことができると思われる。

2.2 メタデータの作成

2.1 で述べたように、作り手の異なるシラバスデータでは、一つの項目値に複数の項目名が用いられているために項目値と項目名の間で一対一の対応をとることができない。このことは、情報の検索や抽出などの処理を行う場合に問題となる。そこで本研究では、多くのシラバスに共通して現れている項目値を用いてメタデータを作成した。以後、このデータを**「共通計画表」**と定義する。共通計画表は各シラバスに共通して現れると予想される項目値それぞれに対して一つの項目名を定めたものである。共通計画表の項目名は、同じ項目値を持つと予想される項目名の集合をもつ。個々のシラバスは項目名と項目値のセットに分解し、項目名を共通計画表の項目名に置き換える。このときの共通計画表の項目名は、もとの項目名を含む集合を持つものが選ばれる。項目名がすべて共通計画表の項目名となるため、個々のシラバスが同じ構造を持つことになり、項目名と項目値は常に一対一の対応を持つ。これにより項目名のばらつきが解消され、検索や抽出などに項目名を利用する事ができる。

具体的に国内の 52 サイトのシラバスを分析し、表 1 に示す共通計画表を作成した。

実シラバスデータを共通計画表に変換する例を示す。表 2 のデータは表 3 のように変換される。対応する項目名がない場合は空値を持つ。

3. シラバスの自動収集

自動収集は、Web 検索システムを利用して幅広く集める段階と、収集したデータに含まれているノイズを除去する段階の二つからなる。

3.1 検索システムによる収集

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

表 1: 共通計画表

シラバス収集の一次段階として、WWW 上のページデータ検索システムを利用した収集を行う。具体的には Google などの WWW 検索システムを利用して自動的にシラバスを収集するプログラムを作成する。このプログラムはシラバスデータを含むページが得られそうなキーワードを用いて検索システムで検索を行い、結果のリストを得る。その後、リストから得られたページを解析し、このページからリンクが張られているページのリストを得る。この時、リンク先の URL がサイト外である場合は無視する。そして、このリストを用いて再びページを得る。また、データを得るファイルは、text ファイル、html ファイルに限定する。このようにして新たに得られたページに対して同様の操作を行い、再帰的にリンク先のページを得ていく。このリンクをたどる作業を、リストから得られたページを基点としてリンクを何段階かたどるまで繰り返す。この結果収集されたページが一次段階の収集データとなる。

3.2 ノイズ除去

一次段階で得られた収集データにはシラバスデータ以外のページが多数含まれている。二時段階では、このノイズの除去を行う。ノイズを除去するために、組織内のリンク構造とシラバスに関連する単語データの出現頻度を用いた各ページの特徴抽出を用いてシラバスデータである部分を推測する。

3.2.2 リンク構造

Web 上で公開されているシラバスデータが単独で存在していることはまれである。通常、シラバスデータはある程度の量がまとまって存在しており、それらのシラバスデータへのリンクをリストとして持つページが存在する。シラバスデータを B、リストのページを A と置くと、図 1 に示すリンク構造をもっていることが多い [6]。このことから、A もしくは B のページを一つでも発見できれば、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。

項目名	項目値
担当者	廣川
科目名	情報基礎演習
内容	計算機についての基礎知識の獲得
教材	１２回で学ぶ情報処理
キーワード	計算機、コンピュータ
授業コード	717
授業学期	１年前期
単位数	2
開講日	水曜 2 限
評価方法	レポート

表 2: 授業計画表（変換前）

項目名	項目値
担当教官	廣川
授業科目名	情報基礎演習
概要	計算機についての基礎知識の獲得
教材	１２回で学ぶ情報処理
キーワード	計算機、コンピュータ
授業コード	717
授業学期	１年前期
単位数	2
曜日と時間	水曜 2 限
評価方法	レポート

表 3: 授業計画表（変換後）

3.2.3 特徴抽出

シラバスに関連する単語データの出現頻度を用いて、各ページについて特徴抽出を行う。シラバスに関連した単語が多数出現するならば、そのページはシラバスデータである可能性が高い。そこで、シラバスに関連する単語の出現頻度の情報を用いた評価関数を用意し、スコア付けを行う。そのページのスコアが一定値以上であれば、シラバスであると判別することができる。シラバスに関連する単語データとしては共通計画表のデータを用いる。この場合、共通計画表が持つすべての単語について出現頻度を調べる手法 (*ev1*) と、項目名を用いてカテゴリ毎に出現頻度を調べる手法 (*ev2*) が考えられる。

4. 実験

プログラムによるシラバスの自動収集を行った。また、3.2.3 で述べた特徴抽出を用いた評価関数をもちい、シラバスであるかの判定を行った。

以下の条件で一次段階の収集を 2001 年 12 月 25 日から 3 日間で行った。

- WWW 検索システムとして Google を利用
- キーワードは「シラバス」

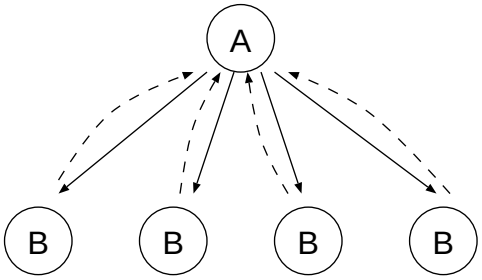

図 1: リンク構造

- リンクをたどる段数は 5

Google から結果として得られた URL は 649 であり、この URL から再帰的に収集を行った結果 452 サイトから 80446 個のファイルデータを収集した。その後、二時段階のノイズ除去を行うために、このファイルデータに対して共通計画表に含まれるすべての単語の出現数をカウントした。また、出現した単語が含まれる項目名をチェックした。この二つの情報を用いてそれぞれのページにスコア付けを行った。スコア付けに使用した関数は以下の二つである。

- $ev1 = \{\sum_{i=1}^{47} w_i\}/47$
  $w_i : i$ 番目の単語が存在するならば 1
- $ev2 = \{\sum_{i=1}^{10} c_i\}/10$
  $c_i : i$ 番目の項目名が存在するならば 1

ev1 は全単語中のうち、存在した単語の割合を示す。*ev2* は全項目名のうち存在した項目名の割合を示す。

5. 評価

4 で述べた関数を用いて、'www.a' で始まる 20 サイト 4281 ファイルに対して評価を行った。まずはこの 20 サイトについて人手で調査を行い、人間がシラバスデータのファイルであると判断したファイルのリストを作成した。このリストを正解集合として、リンクをたどった深さ毎に以下の二つを求めた。

- Hit 率 ：その深さでのシラバスファイル数/その深さの全ファイル数
- Cover 率：その深さでのシラバスファイル数/正解集合のファイル数

その結果を図 2 に示す。深さ 3 以降の Cover 率の上昇は緩やかである。このことから、リンクをたどる数を 3 以上増やしても効果が薄いことがわかった。これは、シラバスデータが 3.2.2 で述べたようなリンク構造をもっているためであると予想できる。そこで次に、深さ 2 のファイル群について評価関数 *ev2* によるスコア付けを行った。その結果を図 3 に示す。この結果から、スコアが 0.4 より大きくなった時点でシラバスデータのファイル数とそれ以外のファイル数が逆転していることがわかる。このことからスコアが 0.4 未満のファイルは、ほとんどのファイルがシラバス以外のファイルであると予測できる。図 4 に、あるスコア未満のファイルを切り捨てた場合について、シラバスデータのファイル数とそれ以外のファイル数、Hit 率、深さ 2 のシラバスファイル全体に対する Cover 率（Cover 率 1)、正解集合に対する Cover 率（Cover 率 2）を示す。スコアが 0.5 をこえると、Cover 率は大幅に下がり、Hit 率の増加幅も小さくなる。このことから、一次段階の収集データに対して、リンクの深さ 2、*ev2* によるスコアが 4 以上という条件を当てはめることで、ノイズを減らすことができるといえる。この場合、Hit 率は 9 割 Cover 率は 8 割となる。

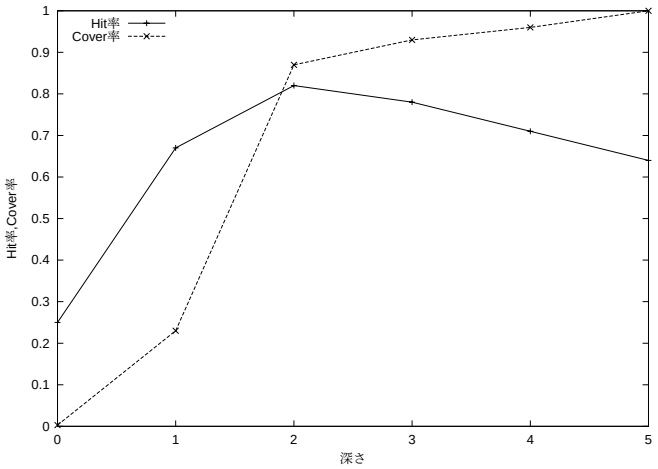

図 2: 深さ毎の評価

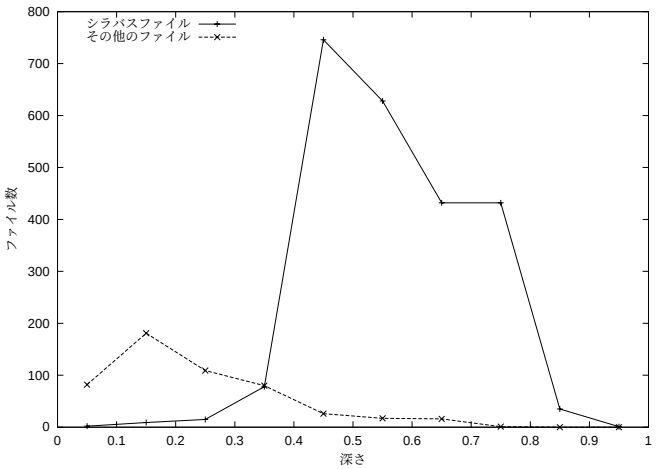

図 3: 深さ 2 の評価

6. 考察

今回の実験では、リンクをたどる深さは 2 ないし 3 で十分であるという結果がでた。この理由としては、シラバスデータが 3.2.2 で述べたようなリンク構造をもっていることがあげられる。収集したページ群をリンク構造に基づいて Web グラフ化したものを図 5 に示す。なお、Web グラフ作成には [7] で示すツールを使用した。この結果からも、シラバスデータは図 1 で示したように、シラバスデータのページ群 (B) とそれらのリストを持つページ（A）よりなるという予測が正しかったとわかる。また、検索システムで得ることのできるページの多くは、B または A のページである。A のページを得たとすると、リンクを 1 段階たどれば B のページ群を得ることができる。B のページを一つ得たとしても 1 段階目で A のページを得ることができ、2 段階目でそのページからそのほかの B のページ群を得ることができる。B のページ

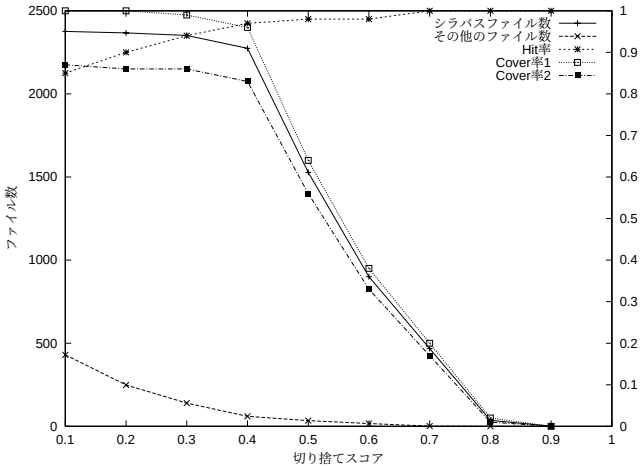

図 4: 深さ 2 の評価 2

を得るために 3 段階以上リンクをたどらなければならない場合とは、A のページが細かく分類されており複数段構成となっている場合か、トップページ等の A や B のページから離れたページであった。しかし、これらの場合も、A や B のページからいちじるしく離れるということは考えにくい。このことから、リンクをたどる数を増やしたとしても、得られるシラバスデータのファイル数は頭打ちになることが予想される。

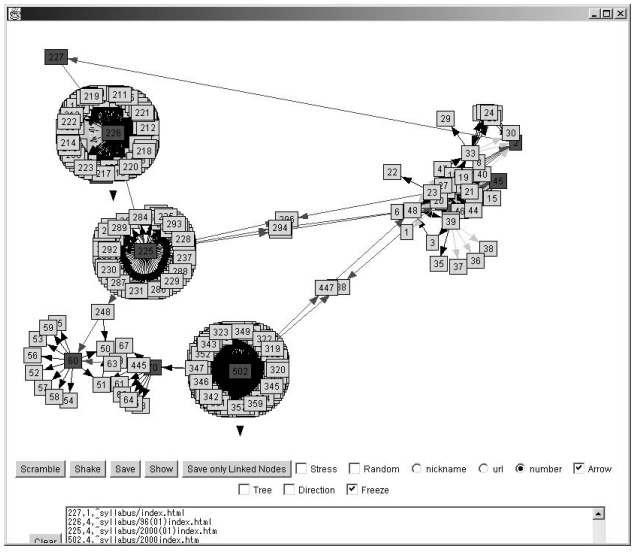

図 5: リンク構造

評価関数による判別では、スコア 0.4 を基準としてシラバスファイルとその他のファイルを分けるのが、Hit 率、Cover 率の両方の面から適しているといえる。スコアが低かったシラバスデータについては、共通計画表作成時のサンプリング数の不足が考えられる。このため、共通計画表を適宜更新していくことが必要となる。また、いちじるしくスコアの低いシラバスデータには項目名が存在せず、表の形となっていなかった。このタイプのシラバスについては、別の手法を考える必要がある。

逆にスコアの高かったノイズファイルについては、入試案内のページや学内のニュースサイトなどが該当した。これらのページには共通計画表に含まれる単語が数多く出現している。本実験で用いた評価関数では、すべての単語が同じ重さをもっているため、このようなページをシラバスデータと区別することができない。そこで、特にシラバスに関連する単語が出現した場合は他の単語が出現した場合よりもスコアを高くする等、単語別に重み付けをする必要がある。

7. おわりに

本論文では、シラバスデータと思われるページを自動的に収集し、その後集めたデータに対して評価を行い、ノイズを除去する手法について考察した。また実際に実験を行い、収集したデータの一部に対して人手による評価と、関数による評価を行った。そして、両者の評価を比較する事で現段階での適当と思われる判断基準を求めた。この判断基準を用いることで、Hit 率 9 割、Cover 率 8 割という値でシラバスを集めることができる。Hit 率、Cover 率については、共通計画表の自動更新や、リンク構造を利用することによってさらに上げることができると考えている。

本論文の手法を応用すると、図 1 の構造をもつページ群を自動で収集できる可能性がある。シラバス以外にこのような構造をもつシラ例としては、日々の新聞記事、不動産情報、地域の病院情報などがあげられる。今後はシラバスに限らず、このような情報を集めて分析するシステムの構築も行いたい。

参考文献

[1] 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)

[2] 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.

[3] 山田泰寛, 池田大輔, 廣川佐千男: “n-gram 交代数を用いた半構造化データの不要部分削除”. 信学技報,Vol.101,No.190,pp.53-60,2001.

[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] 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.

[6] 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.23-35,2002.3

[7] Hirokawa,S.,Taguchi,T.: “KN on ZK - Knowledge Network on Network Note Pad ZK”. Springer LNCS 1532,411-412,1998.