自動収集したWebシラバスデータの分析と考察

A Study and Analysis of Auto Collected Web Syllabus Pages

山田信太郎* 伊東栄典† 廣川佐千男 †
Shintaro Yamada Eisuke Itoh Sachio Hirokawa

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]。多くの大学で教材やシラバスデータの電子化が進み、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。

学習者があるテーマについて勉強したい場合、それに関する講義の受講を考えるが、その際、自分に適した講義を選択するために各教育機関の講義内容を調査する。このような場合、シラバス情報を利用することで、講義内容の概要、教科書や講義資料についての情報を得る事が可能になる。また教育機関の比較も可能になる。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。知識獲得の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. 切り分けられたデータを用いて知識を獲得

現在までに我々はシラバスデータの性質分析を行いメタデータを作成した [2]。また、検索エンジンとリンクを利用した実験用データの収集を行い、その後リンク数と出現単語を用いたフィルタリングを行ってシラバスファイルの判定を行った [3]。

本論文では、フィルタリングを行った前後のファイル数をサイト毎に調査し、ファイル数が大きく減少しているサイト、つまりシラバスファイルの割合の低いサイトについてその理由を考察する。また、これらのサイトについてもシラバスデータを収集できる手法を考察する。

2. データの収集とフィルタリング

データの収集には検索エンジンを利用している。検索エンジンにキーワードを与え、結果として得られるページからリンクを複数回たどることでシラバスデータを収集している。その後フィルタリングを行い、シラバス以外のページの削除を試みている。文献 [3] において収集したデータを分析した結果から、次項に述べるフィルタリングを行っている。

2.1 リンク数によるフィルタリング

分析の結果、3段階リンクをたどることでほとんどのシラバスが得られることが判明した。この理由は、シラバスデータの持つリンク構造にある。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 [4]。実際に収集したデータについて、リンク構造を可視化するツール [5] を用いた結果、この構造をもっていることを確認することができた。検索エンジンで得られるページの多くは、A,Bのページか、それらのページへリンクを張っているページである。このため、リンクを3段階たどれば、ほとんどの場合はシラバスのページへとたどり着くことができる。そこで、リンクをたどる回数を3段階以内としてフィルタリングを行った。

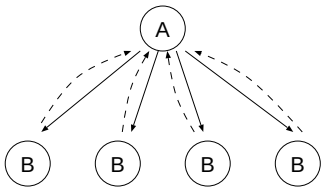

図 1: リンク構造

2.2 出現単語によるフィルタリング

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。分析の結果、ページ内にメタデータの項目のうち4種類以上が出現していればシラバスデータであると判断できることが判明した。そこで、出現する項目数が4種類以上としてフィルタリングを行った。

3. 分析

文献 [3] では、'www.a*' で始まるサイトのファイル群 (これを集合Aとする) について分析した。本論文ではこ

*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

の時に用いたシラバス判定関数 ev を用いて、収集したデータ全体（これを集合 U とする）に対してシラバス判定を行った。

3.1 判定関数 ev

$$ev(d) = \frac{\#\{0 \leq i \leq C | \exists w \in c_i, w \in d\}}{C}$$

w : 共通計画表に含まれる単語

c_i : メタデータの項目

C : メタデータの項目数

ev は共通計画表の全項目名のうち、ドキュメント d のなかに存在した項目名の割合を示す関数である。$ev(d)$ の値が 0.4 以上であれば、シラバスファイルであると判定する。

3.2 フィルタリング後のファイル数

ev により集合 U のうち、6 割のファイルがシラバスファイルであると判定された。フィルタリングを行った前後のファイル数を表 1 に示す。

集合	前	後 (割合%)
U	24128	15423(64)
A	2890	2274(78)

表 1: フィルタリング前後のファイル数

次に各サイト毎にフィルタリング前後のファイル数を調査した。その結果、フィルタリング後のファイル数が激減しているサイトが確認された。これは、シラバスファイルを取得できなかったか、シラバスファイルが極端に少ないことを示している。そこで、これらのサイトについて実際にはどのようなページであったのかを人手によって調査し、その理由を調べた。結果を以下に示す。

1. CGI,JavaScript により管理されている
2. シラバスへ至るまでのリンクが細分化されている
3. シラバスのページからリンクが存在していない
4. シラバスが別のサイトに存在している
5. 個人や研究室のページ（シラバス数が少ない）
6. PDF 形式で公開されている

現在の収集方法では、HTML と TEXT ファイルのみを対象としているため、多数のシラバスを収集し損ねていることが判明した。また、順リンクのみを利用しているため、ページ内にリンクが存在していない場合は収集範囲を広げることができていなかった。リンク先についても同一サイト内に限定したため、別サイトにシラバスが存在している場合は収集することができていなかった。

4. 考察

現在の方法ではうまく収集できないサイトが存在することが判明した。今後はそのようなサイトについても収集できるよう対応させることが必要である。先に触れた原因については、以下の方法が考えられる。

1. CGI のページ
   我々の研究室では、キーワードを用いた検索サイトの特徴抽出技法について研究している [6]。同様の手法を用いることで、CGI による動的データ表示サイトがシラバスサイトであることの判定やシラバスデータの抽出がある程度可能であると思われる。

2. 細分化されたページ
   リンクをたどる深さをサイト毎に変更する。判断基準としては、リンク数やリストになっているか等が考えられる。

3. シラバスのページにリンクが存在しない
   逆リンクを利用することでシラバスページへの到達が可能であると考える。

4. 別サイトに存在
   アンカータグの文字列や、その前後の文から判断する必要がある。

5. 個人のページ
   個人のページを判断するのは困難だが、全体に占める割合が低いため、影響は少ないと思われる。

PDF,JavaScript のページについては対応策を考案中である。

5. おわりに

本論文では自動収集した Web シラバスデータについて判定を行い、シラバスデータを収集できていなかったサイトについてその理由を分析した。また、その対応策についての考察をおこなった。今後はこれらの方法を組み合わせ、より多くのシラバスを収集することができるようにしたい。

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] 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.

[3] 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”, DICOMO 2002.(to appear)

[4] 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.

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

[6] Hirokawa,S., Watanabe,S., Koga,Y., Taguchi,T. : “Automatic Feature Extraction of Search Sites”. Proc. of SSGRR2001.