Webシラバスクローラーの性能評価

松永吉広[†] 山田信太郎[†]
伊東栄典[††] 廣川佐千男[††]

大学等の高等教育機関で体系的に行なわれている講義群の全データを収集，統合，分類できたならば，現在の学問体系の総合目録とよぶことができるであろう．我々は日本国内の大学を対象に，Webで公開されているシラバスデータの収集，統合のためのシステムを開発している．本研究では，シラバスページに現れる特徴的キーワードと，それらの間のリンク情報の特徴を用いることにより、効率的にシラバス・ページ群を収集するクローラーの方式を考案し，実装した．実験的に収集した8万ページのWebページ空間に限定し，収穫率 (HarvestRatio) の観点から，この手法とランダム収集，幅優先収集の比較を行なった．

Efficiency Evaluation of Web Syllabus Crawler

Yoshihiro MATSUNAGA,[†] Shintaro YAMADA,[†]
Eisuke ITOH[††] and Sachio HIROKAWA[††]

1. はじめに

近年，情報技術の発達や情報通信基盤の発達により，ネットワークを活用して，教育や研修を行うといったe-ラーニング4) が盛んに行われている．大学等においても，開講されている講義についての情報であるシラバスデータをWeb上に公開している高等教育機関が増加している．実際のシラバスデータには，講義についての情報である，講義名，教官名，関連科目等の情報を含んでいる．多くの大学等のサイトからシラバスデータを収集し，必要なデータの抽出をすることで，教育データの分析，統合等を行うことができる．このことによって，現在の大学教育についての考察，大学の比較などが，Web上に公開されているシラバスデータ，つまり開講されている講義の情報の観点から行うことが可能になるだろう．

このような背景のもとで，我々はWeb上に公開されているシラバスデータを収集し，それらのデータから抽出，統合を行い，教育データ検索サービス等を提供するシステムの実現を目指している．このシステムの構築について，我々は次のようなフェーズに分けて研究を行っている．

(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シラバスデータからのレコード抽出
(4) 抽出データからの知識獲得

データの収集に関連しては，現在Web空間全体に対して検索サービスを提供しているAltaVista等の一般検索エンジンが必要とするデータではなく，専門検索エンジンなどの特定の分野，領域のデータのみが必要である場合には，一般のWebクローラーを用いて収集を行うと，時間，ディスク，ネットワーク等の資源の多くを無駄にすることになる．このために特定の分野に対象を絞り，効率的に収集を行うと行った研究1),2) が行われている．また，レコード抽出について，Web上に公開されているWebページの多くは，表現方法やレイアウトがWebサイト，Webページごとに異なっているので，容易にWebページの情報を抽出，利用することは不可能である．そこで，このような様々な作者，形式によって書かれたWebページから有用なデータを抽出する技術についても様々な研究5),6) が行われている．このことは，Web上に公開されているシラバスについても同様であり，我々はシラバスファイル中の項目値の抽出等に関する研究3) 等も行っている．

本稿では，Web上に公開されたシラバスデータを効率良く収集するための方法を提案する．この方法の有効性について，ランダムにリンクを辿る方法ならびに幅優先でリンクを辿る方法と比較し，効率を収穫率の観点から評価した．

† 九州大学システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
†† 九州大学情報基盤センター
Computing and Communications Center, Kyushu University

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

2. シラバスのためのメタデータと分析データの収集

2.1 メタデータの作成

我々は，これまでの研究 7) で国内 52 サイトのシラバスデータについての調査を行った．その結果，大学ごと，学部ごとといった Web サイトごとに記述が異なり，体系的な取扱いが困難であることが分かった．そこで，形式の異なる Web 上のシラバスファイルを共通して扱うために，シラバスファイルに共通してみられるであろう特徴的な単語をもとに共通計画表 (**表 1**) を作成した．この共通計画表は，シラバスに関して同義の単語についてまとめたものであり，具体的にはシラバスデータに出現する同義の項目名をそれらを代表する 1 つの項目名で対応させるといったメタデータとなっている．この共通計画表中の単語は，シラバスデータの特徴的な単語であるといえるため，後述するシラバスページ判定にも用いられている．

2.2 実験分析用データの収集，分析

次に，より多くのシラバスデータからその特徴を分析するために Web 上の実際のシラバスデータの収集を行った．具体的には，検索エンジン Google に対し，キーワード「シラバス」で検索を行い，その結果からリンクを抽出，そのリンクを再帰的に辿り収集を行うプログラムを作成し，Web ページを収集した．リンクを辿る際には，同一サイト内を対象に深さ 5 までと制限した．Google の検索結果から 649 の URL を得，これらの URL から，それぞれ上記の制限に基づき再帰的に収集を行った．最終的に，452 サイトから 80446 ファイルを収集することができた．

この 80446 ファイルのうち，ホスト名が “www.a” で始まる 20 サイトの 4273 ファイルを対象として，人手により分析を行った結果，次のことが分かった．4273 ファイル中 2738 ファイルがシラバスデータのファイルであり，シラバスファイルが多く存在しているシラバスサイトには，科目を一覧するリンク集ページと個々の科目についての説明がなされているシラバスページが存在しており，**図 1** のような特徴的なリンク構造をもつものが多いことが分かった．そこで，我々は，図 1 において，個々のシラバスページにリンクしているリンク集ページを A 型，その A 型のページからリンクされているシラバスページのことを B 型と定義した．

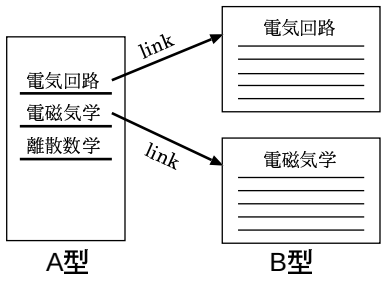

図 1 リンク構造

3. Web シラバスクローラー

B 型のシラバスファイル収集のために前章で述べたシラバスサイトのリンク構造の特徴を用いて，A 型ページからのリンク先の Web ページを収集していけば無作為に収集を行うよりは効率的である．そのためには，そのページが A 型であるかどうかの判定を高い精度で行えることが必要である．そこで，我々は，正例，負例を人手により分けた 4273 ファイルを学習データとして，Web ページを与えられたときに，その Web ページが A 型であるか B 型であるかどうかを自動的に判定する決定木をそれぞれ作成した．この決定木は，その Web ページについての共通計画表中の単語の出現，リンク数の情報から判別を行っていくものである．そして，この作成した決定木を使用した評価実験をおこなった結果 8)，A 型のページについては精度 89%，再現率 87%，また B 型のページについては精度 99%，再現率 99%で判定できた．

次に，この判定関数を Web クローラーの収集の際に用いることで収集の効率性を高めることを考えた．まず，一般な Web クローラーの簡単な流れ図を**図 2** に示す．Web クローラーは収集済/未収集 URL の集合をそれぞれ保持している．まず，未収集である URL の集合から 1 つ取り出す．その URL が Web クローラーによって収集可能であれば，HTTP による要求を行い，Web ページを収集する．次にその収集した Web ページについて，リンクの抽出を行い，適切な URL を既に収集済みでなければ，未収集 URL の集合に加える．これらの処理を繰り返し，Web クローラーは Web ページの収集を行っていく．したがって，特定の分野を対象にする Web クローラーの効率性を考えるときには，未収集リストから次のターゲットの

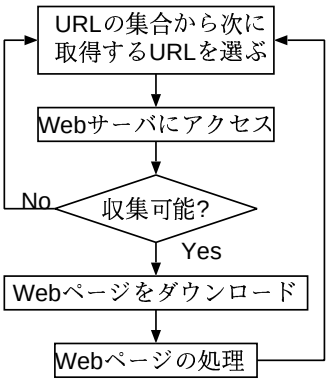

図 2 一般的な Web クローラーの流れ図

選び方，つまり次にどの Web ページを収集するかの決定が，重要となる．

本研究で提案するシラバス収集クローラーでは，次のターゲットを選ぶ際に，A 型と判定されたページからのリンク先の URL を優先的に選び出し収集を行うものである．8)

4. 性能評価

4.1 評価実験

このシラバス収集クローラーの効率性の評価を行った．我々のシラバスファイル収集クローラー (**Atype**) は，前述したように，A 型と判定された Web ページからのリンク先を優先的に収集するといったものである．ここでは，この手法の有効性の確認のために，次に説明する他の Web クローラーを用いても収集実験を行った．まず 1 つ目は，ランダムクローラー (**Random**) である．このクローラーは，未収集 URL の集合から次に収集すべき URL をランダムに選び出し，収集を行うクローラーである．2 つ目は，幅優先探索を行い Web ページを収集するクローラー (**Breadth-First**) である．このクローラーは，未収集 URL の集合を FIFO のキューで扱うものである．したがって，収集を始めたスタート集合からのリンクの深さが浅い URL から徐々に収集していく．また最後は，我々が提案したシラバス収集クローラーと幅優先収集を組み合わせたクローラー (**Atype***) である．このクローラーは，基本的にシラバス収集クローラーと同様に，A 型と判定された Web ページからのリンク先を優先的に収集するものであるが，未収集 URL の集合中にその A 型と判定された Web ページからのリンク先が存在しない場合には，幅優先での収集と同様に，スタート集合からのリンクの深さが浅いものから収集を行うものである．

これらのクローラーを用いて，2.2 章で収集した 80446 ファイルのローカルデータに対して，次の条件で収集実験を行った．まず，Web 空間をこのローカルデータのみに限定し，収集を行うときの初期集合は，この分析用データの収集したときの初期集合，つまり Google の検索結果である 649 の URL とした．Web クローラーはこの初期集合からリンクを辿り収集していく．ただし，収集を許すのは，仮定した Web 空間に存在するファイルのみである．評価としては収穫率，すなはち収集したページ中のシラバスページの割合を用いた．ただし，あるページがシラバスページであるか，否かの判定は，3 章で述べた B 型の自動判定関数を用いた．

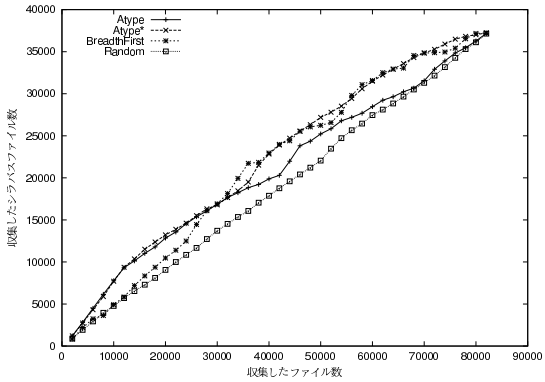

図 3 収集シラバス数

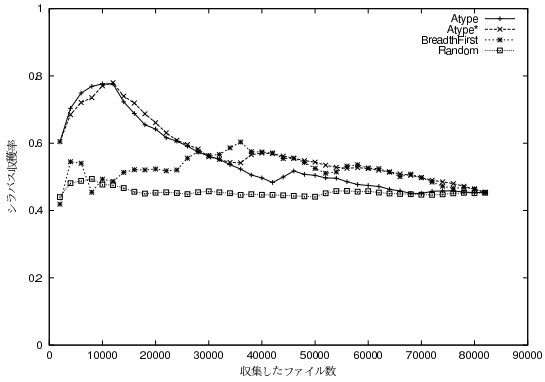

図 4 シラバス収穫率

4.2 実験結果，考察

4 通りのクローラーについて，収集状況の様子は**図 3**，**図 4** のようになった．図 3 の横軸は収集したファイル数，縦軸は B 型ページ判定関数によりシラバスファイルであると判定されたファイル数であり，シラバスファイルの収集の様子を表している．図 4 の横軸も同じく収集したファイル数であり，縦軸は収集したファイル中，シラバスファイルと判定されたファイルの割合である．これは，そのファイル数を収集した時点でのシラバスファイルの精度，つまり収穫率と考えることができる．ただし，この収穫率を計算する上においても B 型のファイルと判定されたものをシラバスファ

イルと見なしている．

図 3 によれば，このスタート集合から収集を行いシラバスファイルを 10000 ファイル程度収集したい場合には，考案したシラバス収集クローラーを用いれば，15000 程度のファイルを収集すればよいが，幅優先探索で収集を行うクローラーでは 20000 程度，ランダムクローラーでは 24000 程度収集する必要がある．また，図 4 からは，仮定した Web 空間から 16000 ファイルを収集した時点では，我々のクローラーは収集したファイルのうち，約 7 割がシラバスファイルであるが，幅優先探索で収集を行うクローラーでは約 5 割，ランダムクローラーでは，約 4 割 5 分しかシラバスファイルを収集できていない．このことから，我々の提案したシラバス収集クローラーが収集を始めてから他の収集法よりも早い段階で多くのシラバスファイルを収集することができており，より効率的であるということができる．

ところが，30000 ファイルを収集した付近では，我々のクローラーと幅優先での収集法と順位が逆転している．これは我々の収集クローラーが A 型ページの判定ミス，または仮定した図 1 のリンク構造をとっていないために，その先にある大量のシラバスファイルを優先的に収集できなかったのに対し，着実に深さの浅い方からもれなく収集を行う幅優先での収集法が有効となったと考えられる．これを改良したのが，Atype* であり，このクローラーではシラバス収集クローラーにおいて，未収集 URL の集合中に，A 型と判定されたページからのリンクがない場合には，幅優先での収集と同様に，深さの浅いページから収集を行う．この収集法は図 4 からいずれの収集法にも大きくは劣っておらず，効率良く収集できていることが分かる．

5. まとめと今後の課題

Web 上にシラバスデータを公開する教育機関が増加しており，それらのデータを統一して利用できるようなシステムは，様々な利用が考えられる．本稿では，そのようなシステムの構築を目指し，シラバスデータを収集するクローラーを提案し，その効率の評価を行った．

シラバスデータを効率の良く収集を行うには，シラバスデータの特徴，リンク構造を分析する必要がある．そのために，シラバスデータの分析を行い，シラバスサイトに存在する「科目を一覧するリンク集ページ」と「個々の科目を説明するページ」について，我々は前者を A 型，後者を B 型と定義した．人手による判定済みデータから作成した決定木を使うことで，高い精度で自動的に A 型，B 型の判定を行うことを可能にした．これらの分析，判定法を利用して効率よくシラバスデータを収集するクローラーを提案し，他のクローラーとともにローカルファイルに対しての収集実験を行った．その結果，我々の提案したシラバス収集クローラーは，他のクローラーに対し，高い精度で収集できており，さらにスタート集合からの深さを考慮に加えることで，より頑健な収集を行えることができることが分かった．

今回の評価実験では，Google の検索結果として得られた 649 個の URL をスタート集合とした．そこで，今後の課題としては例えば各大学の URL のリスト，もしくはその一部をスタート集合としての収集実験，評価を行う必要がある．また今回の収集方法では，A 型の判定の精度に大きく依存している．A 型のページの判定の精度がある程度高いといっても，誤判定を行う可能性は当然ある．したがって，その判定されたページからのリンク先を収集していく段階においても，A 型の判定値を変化させていくような柔軟性を持たせることでよりよく収集できると考えられる．また，今回では A 型のページが見付かったとき，そこからリンクされているページを優先的に辿り収集を行うといったものであったが，今後は何らかの手法を用いて，できるだけ早く A 型のページに辿りつく方法を検討したい．

また，我々はシラバスページから科目名や参考書などの具体的な情報を抽出する研究もおこなっており，これらの他の研究成果と合わせて，統合シラバスシステムを実現していく予定である．

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) S. Chakrabarti, K Punera and M. Subramanyam : “*Accelerated Focused Crawling through Online Relevance Feedback* ”, Proc. WWW2002, 2002.
3) 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 第 57 回 知識ベースシステム研究会, 2002.
4) 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.
5) 古賀康則, 田口剛史, 廣川佐千男 : “検索サイト統合のためのラッパー生成法”, 第 12 回データ工学ワークショップ (CD-ROM), 2000.
6) K. Lerman, C. Knoblock and S. Minton : “*Automatic Data Extraction from Lists and Tables in Web Sources* ”, Proc. ATEM2002, 2002.
7) 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.
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”, エージェント合同シンポジウム (JAWS 2002), pp.371-378, 2002.