Web シラバス情報収集エージェントの試作

山田, 信太郎
九州大学大学院システム情報科学府

松永, 吉広
九州大学大学院システム情報科学府

伊東, 栄典
九州大学情報基盤センター

廣川, 佐千男
九州大学情報基盤センター

出版情報：The transactions of the Institute of Electronics, Information and Communication Engineers. D-I. 86 (8), pp.566-574, 2003-08. 電子情報通信学会情報・システムソサイエティ
バージョン：
権利関係：

論　文

Webシラバス情報収集エージェントの試作

山田信太郎[†]　松永吉広[†]　伊東栄典[††]　廣川佐千男[††]

A study of design for intellingent web syllabus crawling agent

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

あらまし　教育の情報化が進むにつれ，講義内容を紹介するシラバス情報を Web ページとして提示する教育組織が増えている．本研究では，各組織が独自に公開している Web 上のシラバス情報の抽出・統合を行い，ある分野に関する知識を獲得するシステムの実現を目指している．そのためには，シラバスページ収集，レコード抽出，知識提示といった機能を実現する必要がある．本稿では，シラバスデータ収集時により効率よく収集するための方法について考察した．
シラバスサイトには一覧表示するリンク集ページと個々の科目を説明するページが存在している．実際に必要となるのは個々の科目を説明するページであるが，これらのページを効率よく収集するためリンク集ページを利用する．そのために，決定木と重回帰分析を用いてリンク集ページの判定を行った．また，その結果を利用する情報収集エージェントについて考察を行った．

キーワード　ウェブマイニング, 情報検索, 知識ベース, 知的エージェント, クローラー

1. はじめに

情報技術の発達と情報通信基盤の普及に伴い，膨大な情報が電子化され情報ネットワーク，特に Web を介して公開され容易に利用できるようになった．それらは各組織，個人が個別に作成したものであり，関連する内容であっても様々な形式で記述され，系統的な利用は困難である．人類の知識の宝庫とも呼べる Web から，体系的に，あるいは特定の目的にあった情報を収集・分類することは今後の情報社会において最も重要な課題といえる．実際，HTML を代表とする半構造化データから知識を抽出する研究 [3], [15], [16], [19] や，特定のテーマに関する情報を収集する研究 [14] が行なわれている．また，Web データを自動収集するクローラーについても，目的に合致したページだけを効率よく収集する研究がある [1], [2]. 特定の対象について体系的にページ群を収集し，その中から必要とする情報のみを抽出・統合し，用語辞典や百科辞典を構築する試みもある [8].

我々はこのような Web マイニングのテスト・ベッドとして Web で公開されているシラバスを対象とする研究を進めている．国内でも，教材やシラバス等の教育関連情報を電子化し，Web を介して公開する大学等の高等教育機関が増化している．また，組織の評価のためにも社会的にそのような情報公開が強く求められるようになっている．大学等の高等教育機関におけるシラバスは，教育内容のエッセンスと呼ぶことができる．しかも，「本学の教育目的」あるいは「本専攻の教育の特色」というような概念的なものではなく，そこで実施されている講義の実態を伴ったものである．また，高等教育機関の数と分野数を考えると，それらのシラバスを全て収集，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．Web 上のシラバスはその質と量の両面において Web マインニングの重要な課題と考えられる．本研究が目指すシラバス統合システムができれば，特定分野に関する調査や，ある科目についての講義内容の比較や，各組織の教育に対する取り組みの評価などへの利用が可能となり，近年，重要性が高まっている e-Learning [9] の基

[†] 九州大学大学院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.

礎的データにもなると考えている．

Web 上のシラバスからの知識獲得について，我々は以下のようなフェーズに分解し研究を進めている．

（1） シラバスデータの性質分析
（2） 公開されている Web 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

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 [5], [10], [12]．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．これまで，(1) から (3) について実験及び考察を行なってきた [7], [17], [18]．本論文では，(2) のシラバスデータの収集に着目し，効率よくシラバスページを収集する方法と収集したページの精度を向上する方法，そしてこれらを活用する収集エージェントについて考察した．

2. メタデータの作成及び基礎データの収集

2.1 メタデータの作成

[17], [18] においてシラバス統合のために，シラバス項目を表現するメタデータを作成し収集の観点から評価を行なった．

公開されているシラバスページは，多くの場合，一つの科目の説明記述は表の形式になっており，その中の個々の内容は，項目名および項目値のペアになっている．しかし各組織ごとに表の構造も項目名の使い方も異なっている．そこで，項目名の差異を吸収するため，同じ意味を表す複数の項目名をある一つの項目名で代表するメタデータを作成した．このデータを**「共通計画表」**と呼ぶ (表 1)．

実際のシラバスページに記述されている項目名・項目値を共通計画表の形式に当てはめることで，シラバスデータの統合利用が可能になり，検索や抽出などに項目名を利用する事ができる．

また，共通計画表内に現れる単語はシラバスに関連した単語ということができ，シラバスページの判定等に利用することができる．

2.2 基礎データ収集

一般に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を説明するページが存在している．科目を一覧するリンク集ページを A 型，個々の科目を説明するページを B 型とすると，図 1 に示すリンク構造をもっていることが多い [6], [11]．このことから，A 型もしくは B 型のページを発見できれば，リンク構造を利用して残りのシラバスデータも発見できる．

表 1 共通計画表
Table 1 Common syllabus table.

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

この構造を利用し，シラバ分析用の基礎データ収集として，Web 検索システムを利用しページデータを収集した．具体的には，Web サーチエンジン Google を使ったキーワード検索と，その結果からリンクを再帰的にたどり自動的にシラバスを収集するプログラムを作成し，収集を行った．保存するページは TEXT と HTML に限定している．再帰的にリンクを辿る際，同一サイト内へのリンクのみを辿るように制限している．これは，一般には一覧表示するリンク集ページと，個々の科目を説明するページが同一サイト内に固まって存在するためである．

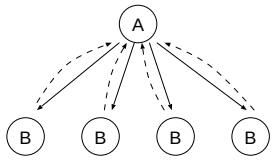

図 1 リンク構造
Fig. 1 Link structure of syllabus site.

以下の条件で基礎データの収集を 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」
- リンクを辿る深さは 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．このデータのうち，'www.a' で始まる 20 サイト 4273 ファイルに対して以下の作業を行った．

表 2　基礎データ (2001 年 12 月 25〜27 日 現在)
Table 2　Test Data (25–27, Dec. 2001.)

ドメイン	総数	B 型	A 型
www.ads.fukushima-u.ac.jp	504	429	4
www.aees.kyushu-u.ac.jp	154	109	3
www.affrs.tuis.ac.jp	273	118	74
www.age.ne.jp	2	2	0
www.agr.kyushu-u.ac.jp	515	493	21
www.agr.niigata-u.ac.jp	164	31	13
www.aj3.yamanashi.ac.jp	27	2	0
www.akeihou-u.ac.jp	114	104	7
www.akjim.yamanashi.ac.jp	1	0	0
www.ams.osakafu-u.ac.jp	2	2	0
www.anan-nct.ac.jp	443	321	10
www.anna.iwate-pu.ac.jp	35	3	0
www.aomori-akenohoshi.ac.jp	192	191	1
www.aomoricgu.ac.jp	228	100	11
www.apc.titech.ac.jp	108	0	1
www.arc.ynu.ac.jp	3	2	0
www.asa.hokkyodai.ac.jp	1148	718	89
www.asafas.kyoto-u.ac.jp	354	109	6
www.asahi-net.or.jp	5	3	1
www.asl.kuee.kyoto-u.ac.jp	1	1	0
合計	4273	2738	241

- 人手により A 型のページリストを作成した
- 人手により B 型のページリストを作成した
- 共通計画表の単語について出現するか調べ，その結果を 0,1 で表す行列を作成した

表 2 に'www.a' で始まる 20 サイトについてサイトごとのファイル数を示す．以上のデータを基礎データとして，実験・評価を行った．

本稿の「共通計画表」は [9] などで援用可能な教育情報のメタデータであり，対応を取ることによって各言語で「共通計画表」に相当するものを構成することが考えられる．また,「syllabus」というキーワードで言語を指定して Google で検索した結果 (2002 年 7 月 17 日現在) でも，英語 1300000 件, 日本語 70100 件, オランダ語 21700 件, フランス語 16300 件, スペイン語 5,320 件等，多くのシラバスページがあることが分かる．従って，各言語での「共通計画表」が構成できれば，本稿で述べた手法は他の言語でも適用できると考えられる．

3. 収集手法の改善

シラバスページを収集し，そのデータから知識を抽出するためには，各科目についての情報が記載されている B 型のページを収集する必要がある．そのため [18] では，特徴的なキーワードの有無を用いて Web 上から B 型のページを判定する事に主点をおいた研究を行ってきた．しかし B 型のページのみに注目していたのでは，効率的な収集が困難であることがわかった．その理由として，B 型のページ間には横の繋がりが少ない点が挙げられる．また，B 型のページはリンク構造の末端であり，それ以上リンクを辿れない場合や，トップページへのリンク等の無駄なリンクのみが存在していることも多い．

一方 A 型のページに注目すると，ページ内に「シラバス」を含む，リンク数が多い，同一サイト内へのリンクが多い，リンク先のページは同一書式であることが多い (B 型のページ) 等，Web 上の全 A 型シラバスページに適用できる特性が存在する．このことから，B 型のページと同じように A 型のページを自動判定できると考えられる．A 型を Web 上から発見できれば，順リンクを辿るだけで，その A 型のページを持つサイト内の B 型ページを漏らさず収集できる．つまり，B 型を 1 つ発見するより，A 型を 1 つ発見する方が，収集されるシラバスページ全体としては効率がよいと考えられる．このような考えから，B 型のページのみを考えるのではなく，A 型のページを発見し，その情報を利用するべきであるとの結論に至った．

3.1　A 型の利点

リンクを辿ることでページ収集を行う場合，集めたい内容と関係の無いページへの無駄なリンクが多数存在する．そのため，より内容が近いページへのリンクを辿ることが出来れば，ページ収集効率が良くなる．2.2 で述べたように，シラバスサイトには A 型のページが存在している．A 型のページを効率良く発見できれば，そこからリンクを辿ることで大量のシラバスを発見できる．リンクを辿って B 型のページを見つける過程には，ほとんどの場合 A 型のページが含まれている．このことから，A 型のページのリンクを優先的に辿ることで収集効率が良くなると考えられる．

また，A 型のページからリンクが張られているページは B 型である可能性が高く，多少評価関数の制限を緩めても問題ないと思われる．しかも，A 型と B 型のページには関連性があるため，A 型のページの情報を用いることでより精度の高い判定を下すこともできると思われる．B 型のページに比べ A 型のページは数が少なく，管理もしやすい．以上のことから，収集手法として B 型のページだけを探すのではなく，A 型のページについても探索を行うことにした．

3.2　A 型の発見方法

A 型のページはシラバスファイルへのリンク集であるため，以下の特徴を持つと考えられる．

- シラバスに関連する単語が出現している
- リンク数が多い
- リンク先が同じディレクトリに集中している
- リンク先の大半は B 型ページである

これらの情報を用いて，A 型ページを判定する評価関数を作成する．

4. ページ型の判定

基礎データの内の'www.a' で始まる 20 サイト 4273 個のファイル（以後集合 S とする）を学習データとして，それぞれのページが何型であるかの判定を行うプログラムを構成した．決定木と重回帰分析の二つの方法について判定プログラムを生成し，評価を行った．

4.1 決定木による判定

学習データに集合 S を用いて，A 型のページを判定する決定木と B 型のページを判定する決定木を作成した．この決定木を用いてページの判定を行い，その結果を評価した．決定木の作成には weka [4] を利用した．

4.1.1 B 型の判定

シラバスに関連する単語の出現データを用いて B 型判定用の決定木を作成した．入力データとして表 3 に示している行列を与えている．“w0”～“w47” は共通計画表に現れる単語を示し 0,1 はその単語がページ中に出現するかを示している．“B?” には B 型のページであれば yes が入る．

表 3 B 型入力データ
Table 3 B-Type input data.

URL	w0	w1	•••	w47	B?
www.ads	0	1		0	no
www.aees	0	0		0	no
:					

集合 S のすべてのデータを用いてこの行列を作成し，“B?” 項目を目的とする決定木を作成した．この決定木を用いて判定を行った結果の混合行列を表 4 に示す．T 行は B 型のファイル数，T 列は B 型であると判定されたファイル数を示している．同様に F 行は B 型以外のファイル数，F 列は B 型でないと判定されたファイル数を示す．精度 99.4%，再現率 99.2%という高い結果を得ることができた．

表 4 決定木による B 型判定
Table 4 B-Type detection by Decision tree.

	T	F
T	2719	21
F	17	1522

次に交差検定 (cross-validation) を行った結果を表 5 に示す．これは，集合 S を 10 個に分割し，そのうちの 9 個を用いて決定木を作成したあと残りの 1 個を評価していったものである．この場合も精度 98.9%，再現率 98.9%と両方とも高い結果を得ることができた．

表 5 決定木による判定 (交差検定)
Table 5 B-Type detection by cross validation.

	T	F
T	2709	31
F	30	1509

以上のことから，決定木を用いることで B 型のページをほぼ正確に判定できると想定できる．

4.1.2 A 型の判定

まずは B 型の時と同様に，シラバスに関連する単語の出現データのみで決定木を作成した．集合 S のすべてのデータを用いてこの行列を作成し，判定を行った結果を表 6 に，交差検定を行った結果を表 7 に示す．T 行は A 型のファイル数，T 列は A 型であると判定されたファイル数を示している．同様に F 行は A 型以外のファイル数，F 列は A 型でないと判定されたファイル数を示す．

表 6 出現単語による A 型判定
Table 6 A-Type detection by term frequency.

	T	F
T	129	112
F	8	4030

表 7 出現単語による A 型判定 (交差判定)
Table 7 A-Type detection by term freqency (cross-validation)

	T	F
T	126	115
F	16	4022

B 型の場合とは異なり，単語の出現データのみで決定木を作成した場合，精度は高い値 (88.7%) を示すが再現率が低い値 (52.3%) となった．そこで，単語の出現データに加えてリンク情報を用いた判定を行った．付加した入力データは，表 8 に示している行列である．B 型判定時に加えて，“L0”～“L2” のリンク情報を付け加えている．“L0” はリンク総数，“L1” は A 型ページと同じディレクトリ内を指しているリンク数，“L2” は同一ディレクトリを指しているリンクの総数を示している． “A?” には A 型のページであれば yes が入る．

集合 S のすべてのデータを用いてこの行列を作成し，判定を行った結果を表 9 に示す．精度 94.6%，再

表 8　A 型入力データ
Table 8　Input data for A-Type detection.

URL	w0	w1	•••	w47	L0	L1	L2	A?
www.ads	0	1		0	20	10	10	yes
www.aees	0	0		0	6	0	3	no
:								

表 9　リンク情報付加後の A 型判定
Table 9　A-Type detection using link structure.

	T	F
T	229	9
F	13	4022

現率 96.2%という高い結果を得ることができた．

次に交差検定を行った結果を表 10 に示す．この場合も精度 89.3%，再現率 87.4%と両方とも 9 割近い結果を得ることができた．

表 10　リンク情報付加後の A 型判定 (交差検定)
Table 10　A-Type detection using link structure (cross-validation).

	T	F
T	208	30
F	25	4010

以上から，単語の出現情報とリンク情報の二つを用いることで，A 型のページについても 9 割近い値で判定ができるといえる．

4.2　重回帰分析による判定

各単語についてページ内にその単語が出現しているかを 0,1 で表す行列を入力として与え，B 型判定をそれらの線形和として求める重回帰分析を行った．単語毎の重みを求め，評価関数の係数として用い，精度，再現率，F 値を求めた．図 2 に結果を示す．決定木を用いた場合と同様，B 型判定は高い値を得た．

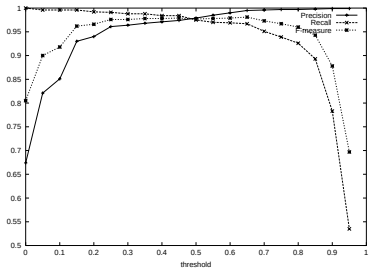

図 2　B 型の判定結果
Fig. 2　Results of B-Type detection

次に単語の出現情報のみを用いた場合の，A 型ページ判定結果のグラフを図 3 に示す．

A 型の判定について，単語の出現情報だけでは決定木の時と同じくよい値を得ることはできなかった．そ

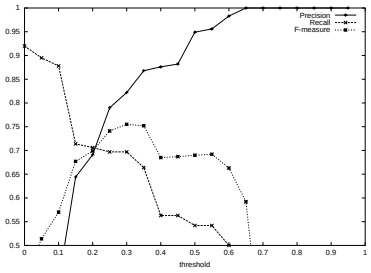

図 3　出現単語による A 型判定
Fig. 3　A-Type detection by term frequency.

こで重回帰分析でも，リンク情報を付け加えた分析を行った．ここではリンク情報として，同一ディレクトリを指しているリンクが 5 つ以上存在するならば 1，5 つ未満であるならば 0 を与えた．その結果のグラフを図 4 に示す．

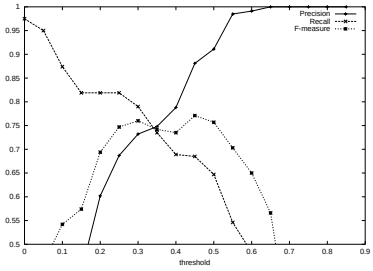

図 4　リンク情報付加後の A 型判定
Fig. 4　A-Type detection using link structure.

リンク情報を用いることで結果が格段によくなった決定木に比べ，重回帰分析では目立った変化は見られず，低い値を示したままであった．これは，リンク情報の扱い方の違いが原因であると考えられる．重回帰分析ではリンクの個数が 5 以上だと 1，未満だと 0 として扱った．これに対して決定木では実数として扱っている．また重回帰分析では線形和として評価値を求めるため，一つの判断基準で判定を行うことになる．決定木では L0,L1,L2 を使って複数の場合分けが行われており，判定に用いるリンクの数も様々である．シラバスのように様々な場合がある対象に対しては，一つの尺度で表現しようとする重回帰分析では対応できないと考えられる．実際に，A 型ファイルには以下の場合がみられた．

- 50 以上の科目の一覧がリンクとしてまとめられている大規模なもの
- 学科や学年ごとに中規模でまとめられたもの
- 5〜6 科目ごとにまとめられた小規模のもの

「リンク数がある値以上」という共通の判定基準で，これら全てを扱うには無理がある．以上の結果，精度・再現率ともに成績の良い決定木による判定を採用した．

5. 情報収集エージェントの試作

前章までの分析結果を用いて，Web シラバス情報収集エージェントの試作を行った．我々の考える Web シラバス情報収集エージェントの目的は，B 型のページを収集することである．このエージェントは，収集を効率よく行うために，あるページが与えられたとき，そのページからさらにリンクを辿るべきかの判断を行う．リンクを辿ってシラバスを効率よく収集するためには，シラバスに関係のあるページをたどっていくことが重要である．シラバスに関係のないページからリンクを辿ってもシラバスにたどり着く可能性は低い．今回試作した情報収集エージェントでは，シラバスに関係するページの判断に A 型の情報を用いた．

A 型のページからリンクが張られているページは B 型である可能性が高い．そこで，A 型のページからリンクが張られているページを優先的に収集する．この情報収集エージェントは次に収集するべき URL のリストを持つが，A 型のページからリンクが張られているページは常に上位におかれる．4 章の結果から，A 型のページ，B 型のページの判定には決定木を用いることにした．

5.1 エージェントの構成

情報収集エージェントの大まかな構成を図 5 に示す．エージェントは大きく分けて 3 つの部分から構成される．

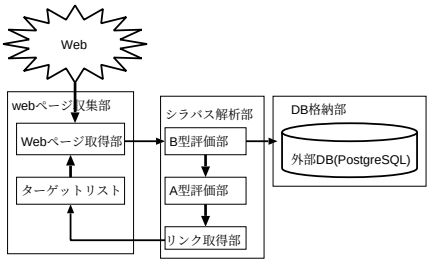

図 5 Web シラバス情報収集エージェント
Fig. 5 Web syllabus crower agent.

Web ページ収集部 この部分は Web 上に存在するページを収集する部分で，ネットワークを直接利用する．ターゲットリストを持ち，この中からスコアの高い順に収集を行う．収集を行う前に，その URL が収集対象であるか，また収集可能であるかを判別する．得られた情報はリンク DB に蓄積される [13].

Web シラバス解析部 この部分は，シラバスに関する解析を行う.「A 型評価部」,「B 型評価部」は取得したページが A 型であるか，B 型であるかの判定を行う.「リンク抽出部」では適切なリンクの抽出を行い，スコアと URL を組にしてターゲットリストへ追加する.

DB 格納部 この部分は，抽出したレコードをデータベースとして蓄積する部分である．データベースのスキーマとしては，表 1 の共通計画表を利用する．また，データの検索にも用いる予定である．

5.2 シラバスページ収集の手段

エージェントには，初期値としてターゲットリストを与える．このリストは，Web 検索エンジンを利用したり，大学の URL リストなどの既知のリストを与える．その後のリンクを辿る手順を以下に示す.

（1） ターゲットリストからスコアの高い URL を取り出し，収集可能か判断する
（2） ページデータを得る
（3） B 型判定を行う
（4） A 型判定を行いスコア付け (0,1) を行う
（5） リンク先の URL を抽出する
（6） 抽出した URL とスコアを組にしてターゲットリストに加える

この手順を図 6 に示す.

ターゲットリストには URL とスコアの組が入っている．スコアはその URL が得られたページが A 型だった場合 1，それ以外の場合は 0 が与えられる．エージェントはスコアが 1 の URL，つまり A 型からリンクが張られているページから収集を行う．収集したページには判定を行う．B 型と判定された場合はデータベースに格納する．B 型と判定されなかった場合は，A 型判定を行った後，リンク先の URL を抽出する．抽出した URL に A 型判定の結果をスコアとして与え，ターゲットリストへと加える．スコア 1 のページがなくなった場合はスコア 0 のページが対象となるが，A 型のページが見つかると，そのページから抽出される URL はスコア 1 となるため A 型のページからリンクが張られているページが優先して収集される．また，シラバスに関係のないページからリンクを辿ってもシラバスにたどり着く可能性は低いため，リンクをたどる回数を 3 回以下に限定している．シラバスデータは同一サイト内に存在することが多いため，サイト外へのリンクも除外している．

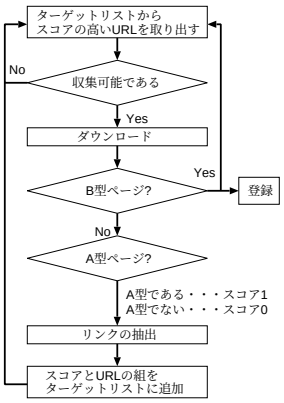

図 6　動作の流れ
Fig. 6　Outline of control flow of the agent.

5.3　収集実験および結果の分析

エージェントの動作の本質は「収集対象の URL リストからリンクを辿り広げる」ことと言える．収集の再現率を上げるため，開始時のターゲットリストをどう選択すべきかも重要な課題である．例えば，全国の大学のトップページ一覧から始めることも考えられるが，その場合，シラバスに関連のないページを多くたどり効率が低下してしまう．本稿では，このように単純なリンク情報だけでなく，一般の検索エンジンも利用する方法で実験を行なった．すなわち初期ターゲットリストとして以下の 2 つを用いて収集を行った．

- Google に「九州大学」「シラバス」の 2 つをキーワードとして与えた結果の上位 100 件
- Google に「九州大学」「シラバス」の 2 つをキーワードとして与えた結果全てから「kyushu-u.ac.jp」のみを抽出

判定に用いる決定木は 4.1.2 の決定木を用いた．使用言語は perl，外部データベースとして PostgreSQL を用いた．12 月 26 日から収集を行い，収集期間は約 1 日であった．

この結果得られたファイル数を表 11 に示す．A 型と B 型のファイル数はエージェントが判定し収集したファイル数であり，全ファイル数はプログラム終了時までに辿った全ページ数である．

表 11　収 集 結 果
Table 11　Results of Crawling.

	A 型	B 型	全ファイル数
上位 100 件	163	2051	11117
kyushu-u 限定	92	1499	5845

次に，A 型，B 型と判定されたファイルについて，人手で検査を行い精度を調べた．結果を表 12 に示す．

表 12　判 定 精 度
Table 12　Precision of type detection.

	A 型	B 型
上位 100 件	0.21	0.81
kyushu-u 限定	0.34	0.93

4.1.2 では高かった A 型の判定精度が極端に落ちている．精度が落ちた理由としては，リンク情報の扱い方が原因だと考えられる．4.1.2 の決定木では，リンク情報のみで判定を行っている部分が存在した．このため，大学の TOP ページや一部の BBS 等のページが A 型と誤判定されていた．基礎データにはこの様なノイズが少なく，学習不十分であった．一方，B 型は高い精度で判定されており，問題なく収集されている．本エージェントの目的は，A 型の情報を用いて効率よく収集を行うことであり，今後は A 型の精度を改善することで，収集効率をあげていく必要がある．

6.　おわりに

シラバスを Web 上に公開する教育機関が増加しており，各組織のシラバスを統一して利用する事で，様々な知識を抽出できると考えている．その様な統合シラバスシステムを実現するために，シラバスデータを収集し，分析するシステムについて研究してきた．

本稿では，Web シラバス情報を収集するエージェントについて考察した．シラバス情報を収集するためには，シラバスサイトの構造や，各ページの特性を分析する必要がある．自動収集する場合には，シラバスではない誤ったページを集めてしまう可能性がある．そこで，収集するデータの精度を向上するために，シラバスページの特性を調査した．

シラバスサイトには,「科目を一覧するリンク集ページ」と「個々の科目を説明するページ」が存在する．前者を A 型，後者を B 型と定義した．既に収集している基礎データを基に，決定木と重回帰分析を用いてそれぞれの A 型あるいは B 型のページの判定をおこなった．また，これらの判定法を利用してより効率よくシラバスページを収集するエージェントについて考察した．また，エージェントを試作して実際にシラバスページを収集し，収集したデータについて分析を行った．その結果，4 章で分析した A 型の判定方法だけでは，不十分であることが判明した．今後はより精度の高い判定方法を考察するとともに，収集効率の点からの評価を行う必要がある．

我々はシラバスページから，科目名や教科書などの具体的な情報を抽出する研究も行なっている [7]．これらの研究成果を組み合わせ，統合シラバスシステムを実現していく予定である．

文　献

[1] C. C. Aggarwal, F. Al-Garawi and P. S. Yu : "*Intelligent Crawling on the World Wide Web with Arbitrary Predicates*", Proc. WWW2001,2001.

[2] S. Chakrabarti, K Punera and M. Subramanyam : "*Accelerated Focused Crawling through Online Relevance Feedback*", Proc. WWW2002, 2002.

[3] 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.

[4] S.J. Cunningham and G. Holmes: "Developing innovative applications of machine learning", Proc. Southeast Asia Regional Computer Confederation Conference, 1999.

[5] 福田賢治, 石野明, 竹田正幸, 松尾文碩 : "極大共通生垣を用いた情報抽出手法の提案", 情報処理学会研究報告 情報学基礎 66-20, pp.151–158, 2002.

[6] S. Hirokawa, T. Taguchi: "KN on ZK - Knowledge Network on Network Note Pad ZK", Springer LNCS 1532, PP. 411-412, 1998.

[7] 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 研究会資料 SIG-KBS-A202,pp.59-64,2002.

[8] 岩爪道昭, 白神謙吾, 畑谷和右, 武田英明, 西田豊明: "オントロジーに基づく広域ネットワークからの情報収集・分類・統合化", 情報処理学会論文誌, 38(3):606-615, 1997.

[9] IEEE : "IEEE P1484 : IEEE Learning Technology Standards Committee (LTSC)", http://ltsc.ieee.org/

[10] 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), 2001.

[11] 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.

[12] K. Lerman, C. Knoblock and S. Minton : "*Automatic Data Extraction from Lists and Tables in Web Sources*", Proc. ATEM2002, 2002.

[13] 松永 吉広, : "Web 空間解析のためのリンクデータベースの設計と実装", 情報処理学会第 65 回全国大会,2003(to appear).

[14] 大槻洋輔, 佐藤理史 : "地域情報ウェブディレクトリの自動編集", 情報処理学会論文誌, 42(9), pp.2310–2318, 2001.

[15] 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.

[16] 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－シリーズ型ＨＴＭＬ文書における類似性の利用－", 人工知能学会誌, Vol.16, No.5, pp.408-416, 2001.

[17] 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.

[18] 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", マルチメディア，分散，協調とモバイル (DICOMO2002) シンポジウム論文集, pp.137-140, 2002.

[19] 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.

（平成 x 年 xx 月 xx 日受付）

山田信太郎

平成 13 年 (2001 年) 九大・工・電気情報工学科卒．同年九州大学大学院システム情報科学府情報工学専攻入学．平成 15 年（2003 年）同修了．主に Web 上の情報収集に関する研究に従事．

松永吉広

平成 14 年 (2002 年) 九大・工・電気情報工学科卒．同年九州大学大学院システム情報科学府情報工学専攻入学，現在に至る．Web 上の情報収集および検索を行うエージェントに関する研究に従事．

伊東栄典

平成 4 年 (1992 年) 九大・工・情報工学科卒．平成 9 年 (1997 年) 九州大学大学院システム情報科学研究院博士後期課程修了．同年 4 月，九州大学大型計算機センター助手．平成 12 年 (2000 年)10 月，九州大学情報基盤センター助教授，現在に至る．主に自律分散エージェント，分散システム，Web マイニング，ソフトウェア工学に関する研究に従事．情報処理学会，人工知能学会 各会員．博士 (情報科学)．

廣川佐千男　（正員）

昭 52 九大・理・数学卒，昭 54 同大学院理学研究科修士課程了．静岡大学工学部情報工学科助手，九州大学教養部教養部助教授，九州大学理学部助教授，九州大学大学院システム情報科学研究科教授を経て，現在，九州大学情報基盤センター教授．リンク情報による WWW 空間の解析，半構造化データからのデータ・マイニング，インターネット・ナビゲーション，および型理論の研究に従事．情報処理学会，電子情報通信学会，人工知能学会，ソフトウェア科学会，数学会，ASL，EATCS 各会員．博士（理学）．