大学評価・学位研究　第１号　平成17年３月（論文）
［独立行政法人大学評価・学位授与機構］

ウェブ文書からの情報抽出に関する研究の概観
―シラバスデータへの適用に向けて―

A Survey on Methods for Extracting Information from Web Documents: Considering their Application to Syllabus Data

芳鐘　冬樹，井田　正明，野澤　孝之，宮崎　和光，喜多　一
YOSHIKANE Fuyuki, IDA Masaaki, NOZAWA Takayuki, MIYAZAKI Kazuteru and KITA Hajime

Research on Academic Degrees and University Evaluation, No. 1 (March, 2005) [the article]
National Institution for Academic Degrees and University Evaluation

ウェブ文書からの情報抽出に関する研究の概観
―シラバスデータへの適用に向けて―

芳鐘　冬樹*，井田　正明**，野澤　孝之*，宮崎　和光***，喜多　一****

１．はじめに

近年，インターネット，特にWWWの普及が進み，膨大な情報がウェブページを介して発信されている[1]。大学からも様々な情報が学内外に発信されており，ネットワークを介してシラバス（授業計画）を公開する大学も増えてきている。シラバスは，授業の選択と履修時のガイドのための学生に対する情報提供を本来の目的として作成・配布されているが，授業内容を最も詳しく示す資料であるため，収集したシラバスから，各項目を適切に抽出・編集できれば，大学ごとの傾向の分析や，科目の関連性の分析など，教育課程の構造解析への活用，さらには，それに基づく大学評価の支援への活用も期待できる。

ただし，ウェブページ上の文書（以下，ウェブ文書と呼ぶ）には，含まれる情報の内容や質が多様であるだけでなく，そのフォーマット（HTMLなどの構造）もまた多様であるという問題がある。ウェブページ上で公開されているシラバスも，その点は同じであり，項目の並び順や用いられるHTMLタグなどは，大学によって（さらには同じ大学でも学科によって）様々である。そのように「不均一な」ウェブ文書から必要な情報を適切に抽出するための方法として，いくつもの手法が提案されている。本稿では，それらウェブ文書からの情報抽出に関する研究について整理・概観したうえで[2]，シラバスデータへの適用可能性について議論する。

２．抽出手法の概観

ウェブ文書からの情報抽出の手法は，抽出の際に手掛かりとする情報から，次の２つに大きく分類できる。

⑴　抽出対象，あるいはその前後の**言語表現の特徴**を手掛かりに抽出する手法
⑵　文書を**構造化データ**として扱い，抽出対象の**文書中の位置**（前後のHTMLタグ，出現順序，木構造中の位置）を手掛かりに抽出する手法

また，それぞれ，「手掛かり」をデータ観察に基づき人手で特定する手法と，機械学習などで自動的（あるいは半自動的）に特定する手法とがある。提案されている手法のほとんどは，⑴⑵のいずれか，あるいは両者を組み合わせたものである。以下，⑴言語表現に基づく抽出，⑵構造情報に基づく抽出のそれぞれについて，代表的な研究を紹介する[3]。

*　独立行政法人大学評価・学位授与機構　評価研究部　助手
**　独立行政法人大学評価・学位授与機構　評価研究部　助教授
***　独立行政法人大学評価・学位授与機構　学位審査研究部　助教授
****　京都大学　学術情報メディアセンター　教授

[1] 例えば，検索エンジンGoogle（http://www.google.co.jp/，http://www.google.com/）がインデクシングしているウェブページは，2004年３月31日現在で4,285,199,774ページにものぼるが，１つの検索エンジンがカバーするページは実際に存在するページの約１／３にすぎないとも言われ（Lawrence & Giles，1999），インターネット上にいかに膨大な量の情報が溢れているかが分かる。

[2] 本稿では，HTML形式のウェブ文書からの情報抽出手法を中心に整理する。シラバスの場合，PDF形式など他の形式で提供されているものも多いが，本稿で紹介する手法の大部分は，若干の調整で他の形式にも適用可能である。

[3] ウェブ文書からの情報抽出に関する研究は，非常に多く存在しているため，ここでは代表的な研究を紹介するにとどめる。また，Visual Web Task（http://www.lencom.com/VisualWTSite.html）など，情報抽出のためのソフトウェアも存在していることも付記しておく。

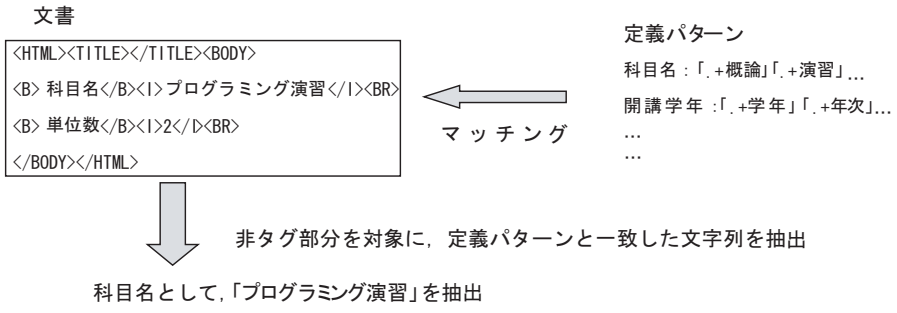

図1　文字列のパターンマッチングによる抽出

2.1　言語表現に基づく抽出

抽出対象，あるいはその前後の言語表現の特徴を手掛かりとする場合，フィールド（HTMLタグに囲まれた非タグ文字列部分）を切り出し，それらを抽出源としたうえで，技術的には，構造化されていないプレーンテキストからの情報抽出とほぼ同じ手法が適用できる。具体的には，A．文字列のパターンマッチング，B．辞書情報の利用，C．テキストの類似度に基づく分類，といった手法がある。

A．文字列のパターンマッチング

予め，抽出対象が含む特徴的な（語や字種などの）パターンを定義しておき，それと一致する文字列を判別して抽出する手法を用いた研究は多い。例えば，井出ら（1996）は，製品情報中の価格・発売日を，佐藤（2001）は，住所情報中の郵便番号・電話番号を対象に，それらに対応する正規表現パターンを用いて抽出している。また，抽出対象の語尾などに現れる特徴的な表現（シラバスの科目名を例に取れば，語尾の「概論」「演習」など）を用いて抽出しているものとしては，佐藤ら（1994；1995），渡辺ら（2004）などがある。

抽出対象そのものではなく，その前後に現れる言語表現に着目する手法もある。特に，表形式の文書の場合，項目名（記述の種類・属性）と項目値（それに対応する具体的な記述，つまり抽出対象）が対になっている場合が多いため，項目名を判別できれば，それに続く項目値の抽出も可能になる。対応する項目名を特定して抽出を行った研究には，見舘＆佐藤（1999），伊東ら（2002ａ；2002ｂ；2003）などがある。伊東らは，同じ項目でも，ページによって，項目名の表記が異なる場合も多いことを考慮して，1つの項目に対して複数の項目名を用いている（例えば，シラバスの科目名では，「授業科目名」「授業科目」「テーマ」「研究主題」「講義科目」「科目名」）。

フィールド（タグに囲まれた非タグ文字列部分）全体ではなく，その一部分が抽出対象となるような場合には，項目名に限らず，抽出対象の前後に現れる言語表現が，抽出箇所特定のための重要な手掛かりになる。抽出対象の前後に現れる固定パターンを利用して，情報抽出を行った研究には，井出ら（1996），佐藤（2001）などがある。井出らの手法では，出現頻度と隣接文字のエントロピーに基づき（文字列のまとまりの計測に基づき），学習用の文書集合から固定パターンを自動的に取り出したうえで，それらの固定パターンを利用して抽出用のテンプレート[4]を作成している。例えば，「.+に販売を始める」というテンプレートが作成された場合，「.+」の箇所を発売日として文書から抽出する。

図1に，文字列のパターンマッチングによる抽出手法の簡略な概念図を示しておく。

B．辞書情報の利用

文字列のパターンマッチングによる抽出の延長であるが，マッチングに辞書を用いた研究もある。抽出対象を列挙した辞書（情報源）が存在する場合，この手法は有用である。佐藤（2001），住所辞書（郵政省が公開した郵便番号のデータベースを加工して作成）をもとに，ウェブ文書から住所の抽出を行っている。大槻＆佐藤（2000）は，地域情報の自動収集・編集を目的にして，ウェブ文書中の都道府県名・地域名の出現箇所をマッチングで特定し，それをアンカ文字列とするURL

[4] 本稿では，「抽出の際に鋳型となる共通の枠組み」という意味でテンプレートという語を用いている。

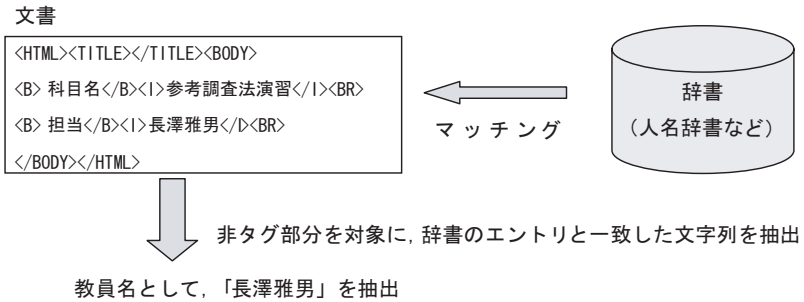

図 2　辞書情報を利用した抽出

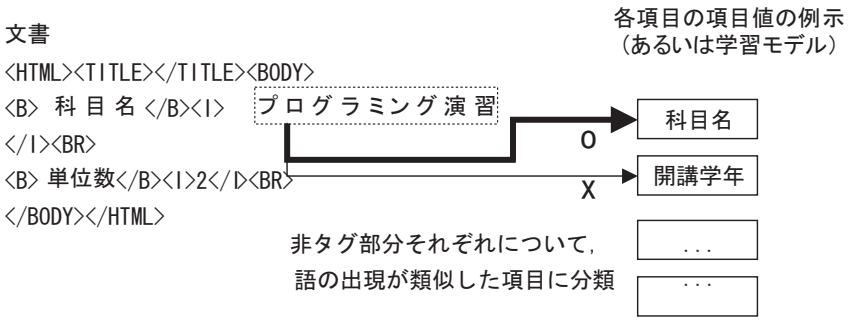

図 3　テキストの類似度に基づく分類

（つまり，その地域に対応するリンク先ページ）の抽出を行っている。また，渡辺ら（2004）は，形態素解析ツール茶筌（松本ら，2003）の固有名詞辞書に基づいて（茶筌の品詞判定に従って），シラバスのページから教員名の抽出を行っている。

適当な既製の辞書が存在しない場合でも，抽出対象の表現がある程度限られていれば，抽出対象を列挙して辞書を自作することが可能である。井出ら（1996），冨田ら（1996b）は，製品・商品情報における製品種別や販売元などについて，サンプルをもとにするなどして辞書を作成し，それを抽出に利用する手法を提案している。

図 2 に，辞書情報を利用した抽出手法の簡略な概念図を示しておく。

C．テキストの類似度に基づく分類

抽出候補がある項目に対応するかどうかを，定義したパターンに一致するか否かで判別するのではなく，各項目の例示や学習モデルに基づいて，抽出候補を類似した項目に分類する手法もある。特に，抽出対象が単語や名詞句などではなく，ある程度の長さを持つ文章である場合は，有効なパターンの作成が困難なため，単純なパターンマッチングよりもそのような手法が有用と考えられる。

白田（2001）は，抽出候補（非タグ文字列部分）それぞれについて，予め作成しておいた各項目の例示文章との類似度を計算し，類似度に基づいてどの項目に該当するかの判別を行う手法を提案している（白田が扱った対象は，商品（洋服）情報におけるサイズ説明部とデザイン説明部）。類似度の計算には，語の出現頻度分布の類似度や DP マッチングによる文字列類似度などが利用できる。

また，項目の判別に，SVM（Support Vector Machine）によるテキスト分類を応用した研究もある。板井ら（2003）は，シラバスのページから，科目名（和），科目名（英），教官名，学年，学期，単位数の 6 項目を抽出するのに，SVM を適用している。具体的には，6 項目それぞれを分類クラスとし，各クラスについて，学習用の文書集合に基づいて，2 値分類（one-versus-rest）の学習モデルを作成し，それらを用いて抽出候補のクラスへの分類を試みている。

図 3 に，テキストの類似度に基づく抽出手法の簡略な概念図を示しておく。

2.2　構造情報に基づく抽出

抽出対象あるいはその周辺の言語表現，つまり非タグ部分の特徴を手掛かりにするのではなく，タグ部分のパターンや，タグによって示される構造の中での位置を抽出に利用する手法もある。そのような手法を，A．抽出対象の前後の HTML タグを特定する抽出手法と，B．文書の大域構造

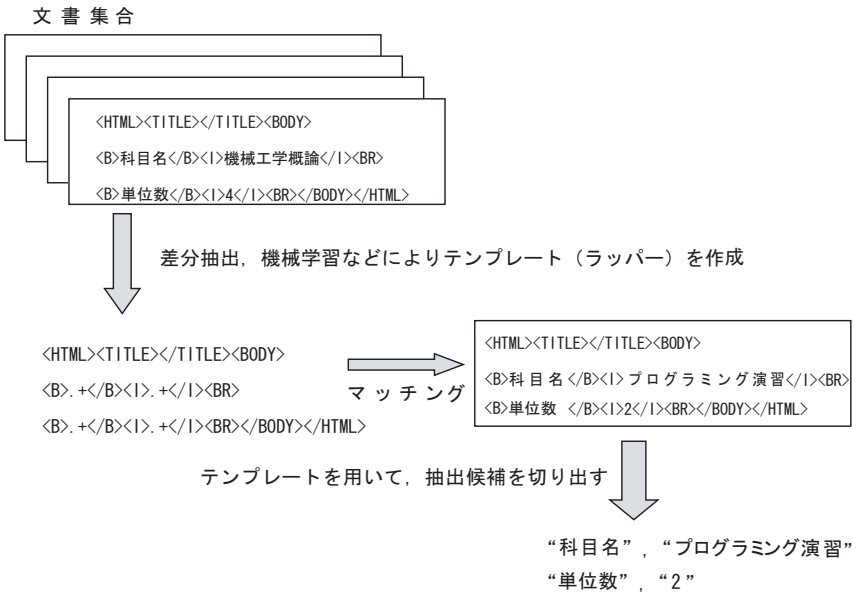

図4 前後のHTMLタグの一定性に基づく抽出

の推定に基づく抽出手法とに分けて整理する。

A．抽出対象の前後のHTMLタグを特定

タグパターンの一定性・類似性を抽出に用いる手法は，「抽出対象を挟むタグ文字列は，特徴的なパターンとして表現される」という考え方を背景に持つ。タグパターンを抽出に用いる手法のうち，パターンの特徴を人手で記述している研究としては，佐藤ら（1994；1995）などがある。佐藤らは，会議情報の抽出において，ヒューリスティクスに基づいた項目ごとのスタイル情報（センタリングや箇条書きなど）の特徴を，言語表現の特徴を補う形で利用している。

それに対して，抽出対象の前後のタグパターンを（半）自動的に特定する手法としては，冨田ら（1998a；1998b），伊東ら（2002a；2002b）などが用いている複数文書の差分に基づくテンプレート作成がある。同一サイト内の同じタイプの文書（例えば，同じ大学・学部のシラバスのページ）は，同一のソフト（ワード・プロセッサやホームページ作成ソフト）を使用して作成されている，あるいは同一のデータベースなどから自動的に生成されていることが多く，そのような場合，通常，HTMLの記述スタイル（タグパターンや，項目が並ぶ順序）は文書によらず共通している。したがって，複数の文書を比較して，共通のタグパターンを除いた差分を，抽出候補と見なすことが可能である。冨田らや伊東らは，差分を取る（差分を正規表現に置き換える）ことで，複数の文書が共通して持つタグパターンのテンプレートを作成し，それと処理対象文書とをマッチングすることによって，目的とする情報の抽出を行っている。切り出す項目値（抽出候補）と項目名の対応付けについては，前節で述べた言語表現の特徴を利用する手法と，タグパターンの訓練例をもとに機械学習を行う手法とがある。機械学習によるラッパー（テンプレート）の発見に関しては，池田ら（2002），山田ら（2004）に詳しい解説がある。

図4に，前後のHTMLタグの一定性に基づく抽出手法の簡略な概念図を示しておく。

B．文書の大域構造を推定

特定の項目の前後のタグに注目するというよりも，より大域的な表構造や木構造の推定を通して抽出を行う手法もある。例えば，嶋田＆遠藤（1999）は，複数の列・行を記述するHTMLタグ（TDタグのcolspan属性およびrowspan属性）の解析を行い，抽出する項目値と項目名の対応付けを行っている。また，伊東ら（2002a）は，項目名と対応する項目値が隣接せず，項目名がいくつか繰り返した後に，項目値が同じ数だけ繰り返す構造を持つケース（例えば，「学年・学期，4年・前期」）を考慮した項目名と項目値の対応付けを行っている。

機械学習により，HTML 文書の木構造の推定を行った研究としては，村上ら（2001）などがある。村上らは，抽出対象となる項目値（ノード）に至る木構造の中の経路（パス）をタグ構造で表現するツリー・ラッパーの学習アルゴリズム（訓練例に基づいて抽出パスを計算するアルゴリズム）を提案している。学習によって得たラッパーを処理対象文書に適用することで，抽出パスに一致するパスを探して，一致したノードを抽出対象として取り出すことができる。

抽出対象とする各項目の，文書構造の中での並び順に着目した研究としては，板井ら（2003），渡辺ら（2004）がある。渡辺らは，言語表現の特徴の文書横断的な分析に基づき，抽出対象が現れる構造上の位置を推定して，それを抽出に利用している。一方，板井らは，各項目が出現する順序を，HMM（隠れマルコフモデル）に基づいて学習する手法を提案している。板井らの手法では，項目名および項目値の種類それぞれを１つの状態と考え，学習用の文書集合から，各状態間の遷移確率を計算して HMM の状態遷移行列を作成している。そして，その学習モデルをもとに，ビタビアルゴリズムで確率積が最大になる最尤状態遷移系列を探索し，処理対象文書における項目の並び順を推定している。

3．シラバスデータからの情報抽出

本章では，ウェブページ上のシラバスの特徴について概略を述べた後で，シラバスデータを対象とする先行研究を紹介し，シラバスデータからの情報抽出の手法について考察する。

3.1　シラバスデータの特徴

Ａ．項目の多様性

冊子体のシラバスと同様，ウェブページ上で公開されているシラバスも，授業内容を示すための項目を多数含んでいる。井田ら（2003；2004）は，シラバスの XML データベース化を目的に，シラバスに含まれる項目について整理を行っており[5]，一般的な項目名として「授業コード」「科目名」「開講対象学年」「開講学期」「曜日」「時限」「単位数」「授業形式」「教室」「教官情報」「授業概要」「授業計画」「履修により達成される目標」「成績評価方法」「教科書」「参考書」などを抽出している。ただし，すべてのシラバスが，これらの項目のすべてを備えているわけではなく，大学や学部によって項目は異なっている。

上に挙げたようなシラバスの項目には，項目値の記述が長いもの，短いもの，表現のバリエーションが大きいもの，小さいものなど，様々なタイプのものがある。例えば，

⑴「開講対象学年」「開講学期」「曜日」「時限」「単位数」「授業形式」など，取りうる値のバリエーションが（表記のゆれはあっても実質的には）かなり小さく，限定的な項目。
⑵「科目名」「教官情報（教官名）」など，記述が短く，また記述全体としては表現のバリエーションが大きいものの，記述の一部に注目すると，ある程度特定が可能な項目（「科目名」の例では，「実験」「概論」「演習」などの語尾，「教官情報（教官名）」の例では，姓の部分）。
⑶「授業概要」「授業計画」「履修により達成される目標」など，記述が比較的長い項目。

上記のように分類することができる。言語表現に基づいて，シラバスデータから情報抽出を行う際には，抽出の対象とする項目のタイプを考慮に入れる必要がある。

Ｂ．構造の多様性

渡辺ら（2004）は，ウェブページ上のシラバスを対象に，項目の出現順序の調査を行い，

⑴　ページの先頭部に「科目名」や「科目英文名」が現れる。
⑵　ページの中間部に「開講学年」「曜日」「時限」「単位数」「教員名」などといった科目の基本事項が記述される。
⑶　ページの末尾部に「授業計画」「教科書」「成

[5] シラバス（教育情報）を構造化表現した井田らの XML Schema は，http://svrrd2.niad.ac.jp/syllabus/EDB10.xsd にて公開されている。今後の展望として，例えば，このような共通フォーマットが普及し，各大学がそれに則ったデータを作成・公開するようになれば，シラバスデータからの情報抽出，そしてそれに基づくシラバス分析が容易になるものと考えられる。

績評価方法」などが記述される。

という傾向を報告している。ただし，上で述べたとおり，大学や学部によって，実際に記述されている項目はまちまちである。また，およその傾向として，⑴⑵⑶のようなことは言えても，実際に項目が現れる順序は，やはり大学・学部によって異なっている。さらには，各回の「授業計画」など，複数回繰り返される項目もあり，情報抽出を行う際には，こうした構造の多様性（用いられるHTMLタグの多様性も含めて）を考慮しなければならない。

3.2 抽出手法のシラバスデータへの適用

前章では，ウェブ文書からの情報抽出一般について，先行研究の整理を行った。それらの研究が対象とするデータは様々であり，既に触れたようにシラバスデータを扱った研究も存在している。

伊東ら（2002a；2002b；2003）は，複数文書の比較に基づく抽出テンプレートの作成，言語表現の特徴に基づく項目の種類の推定という手法で，「担当教官」「授業科目名」「概要」「教材」「関連科目」「キーワード」「授業コード」「授業学期」「単位数」「曜日と時間」「評価方法」といった項目の抽出を行っている。板井ら（2003）は，言語表現の特徴に基づくSVMによるクラス分類と，HMMによる表構造推定に基づく分類の組合せで，「科目名（和）」「科目名（英）」「教官名」「学年」「学期」「単位数」の6項目の抽出を行っている。また，渡辺ら（2004）は，抽出対象の言語表現の特徴と，抽出対象が現れる構造上の位置をもとに，「科目名」「科目英文名」「開講学年」「開講学期」「曜日」「時限」「必修等」「単位数」「教官名」の9項目の抽出を行っている。抽出実験の結果，板井ら，渡辺ら，ともに，およそ90%以上の精度・再現率を得ている。（板井ら（2003）は，精度の評価のみを示している。伊東ら（2002a；2002b；2003）には，抽出パフォーマンスは示されていない。）

伊東ら，板井ら，渡辺ら，すべて，言語情報の利用と構造情報の利用を組み合わせた手法になっている。言語情報を利用した手法は，抽出候補そのものの特徴だけで判別できるため，文書の構造に依存しない抽出が可能，つまり構造のゆれに対してロバストという長所を持つが，抽出対象のタイプに依存する（例えば，記述が短く，かつ表現のバリエーションが大きい項目は，言語表現のパターンの特定が難しく，正確な抽出が困難である）という短所もある。一方，構造情報を利用した手法は，記述の表現のゆれに対してロバストであるものの，処理対象とする文書集合の構造がおよそ一定でないと，正確に抽出することは難しい。

シラバスの場合，言語表現の固定パターンの作成が容易な項目（開講学年・学期，単位数など）が多いといった点では，言語情報の利用が有効であり，また，同じ大学・学部では，文書の構造が概ね一定であるため，構造情報の利用もある程度有効と言える。したがって，伊東ら，板井ら，渡辺らの手法のような，言語情報の利用と構造情報の利用との組合せは非常に有効であると考えられる。ただし，前述のとおり，シラバスに含まれる項目には，様々なタイプのものがあり，抽出対象とする項目の特性をも考慮して，有効な手法の組合せを考えなければならない。例えば，「授業概要」「授業計画」「履修により達成される目標」など，記述が比較的長い項目の判別は，先行研究ではまだ試みられていない[6]が，それらの判別には，構造情報を利用するとともに，SVMなどテキスト分類の技術を応用することも有効だろう。

最後に，シラバスデータからの情報抽出において有効と考えられる手法を，抽出対象とする項目のタイプ別にまとめておく。

⑴ 「開講対象学年」「開講学期」「曜日」「時限」「単位数」「授業形式」など，取りうる値のバリエーションが（表記のゆれはあっても実質的には）かなり小さく，限定的な項目。
→ 文字列のパターンマッチング

⑵ 「科目名」「教官情報（教官名）」など，記述が短く，また記述全体としては表現のバリエーションが大きいものの，記述の一部に注目すると，ある程度特定が可能な項目。
→ 文字列のパターンマッチング，辞書情報の利

[6] 伊東らは，概要，計画，目標などの項目を「概要」として一括りで抽出しており，それらの中での区別はしていない。

用

⑶「授業概要」「授業計画」「履修により達成される目標」など，記述が比較的長い項目。
→ テキストの類似度に基づく分類，抽出対象の前後の HTML タグを特定，文書の大域構造を推定

4．おわりに

本稿では，ウェブ文書からの情報抽出手法を，言語情報を利用する手法と構造情報を利用する手法に分けて整理し，また，それらの手法のシラバスデータへの適用について考察を行った。本稿は，シラバスから各項目の抽出を行う部分に焦点を当てて議論しているが，当然，その前後のプロセス，すなわちシラバスのページの収集と，抽出した情報の分析も重要である。シラバスのページの自動収集については，山田ら（2002a；2002b；2002c）などの試みがある。また，シラバスから抽出した情報を活用した研究としては，科目分類支援への適用を行った宮崎ら（2003）の試みがあるが，他にも，教育課程の大学間の比較や，科目の関連性などに関する知識発見への活用も可能と考えられ，それらの方面でも分析手法の確立が期待される。

謝　辞

本稿をまとめるにあたりご協力いただいた「大学評価情報の構造解析と評価プロセスへの応用の研究会」参加者の皆様に謝意を表します。

参考文献

1）井田正明，宮崎和光，芳鐘冬樹，喜多一（2003）シラバス XML データベースシステム構築に関する考察．情報処理学会第65回全国大会，２Ａ－６，p.４－247－４－248.
2）井田正明，芳鐘冬樹，野澤孝之，宮崎和光，喜多一（2004）シラバス XML データベースシステムの試作．情報処理学会第66回全国大会，３Ｃ－５，p.４－373－４－374.
3）井出裕二，藤吉誠，永井秀利，中村貞吾，野村浩郷（1996）テンプレートを用いた新聞記事からの製品情報抽出．システム情報処理学会 NL 研究会，96，NL 115－12，p.83－90.
4）池田大輔，坂本比呂志，有村博紀（2002）ウェブデータマイニング．システム／制御／情報，**46**（４），p.177－183.
5）板井久美，高須淳宏，安達淳（2003）HTML からの情報抽出と統合．NII Journal，**6**，p.９－19.
6）伊東栄典，松永吉広，山田信太郎，廣川佐千男（2003）Web シラバスからの DB 構成．第17回人工知能学会全国大会，１Ｄ４－08.
7）伊東栄典，山田信太郎，廣川佐千男（2002a）Web シラバス統合のためのレコード解析．人工知能学会研究会資料，SIG-SWO-A201，p.（05－１）－（05－７）.
8）伊東栄典，山田信太郎，松永吉広，廣川佐千男（2002b）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（第57回知識ベースシステム研究会 SIG-KBS），SGI-KBS-A202，p.59－64.
9）Lawrence, Steve and Giles, C. Lee (1999) Accessibility of information on the web. *Nature*, **400**, p.107－109.
10）松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸（2003）『形態素解析システム『茶筌』version 2.3.2使用説明書』，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座，18p.
11）見舘潔，佐藤理史（1999）教官公募情報のダイジェスト自動生成．情報処理学会第58回（平成11年前期）全国大会講演論文集，**3**，p.87－88，４Ｔ－１.
12）宮崎和光，井田正明，芳鐘冬樹，喜多一（2003）電子化されたシラバスに基づく科目分類支援システムの開発について．第２回情報科学技術フォーラム，p.381－382.
13）村上義継，坂本比呂志，有村博紀，有川節夫（2001）HTML からのテキストの自動切り出しアルゴリズムと実装．トランザクション「数理モデル化と応用」，**42**，SIG14，p.21－24.
14）大槻洋輔，佐藤理史（2000）ワールドワイドウェブを知識源とした地域情報の自動編集．情報処理学会研究報告，ICS，知能と複雑系，**3**，p.165－172.
15）佐藤円，佐藤理史，篠田陽一（1994）電子ニュースにおけるダイジェスト機構の実現．情報処

理学会第49回（平成6年後期）全国大会講演論文集，**3**，p.211－212，3K－3．
16）佐藤円，佐藤理史，篠田陽一（1995）電子ニュースのダイジェスト自動生成．情報処理学会論文誌，**36**（10），p.2371－2379．
17）佐藤理史（2001）ワールドワイドウェブを利用した住所探索．情報処理学会論文誌，**42**（1），p.59－67．
18）嶋田和孝，遠藤勉（1999）製品性能表からの特徴データの抽出．自然言語処理，133－15，p.107－113．
19）白田由香利（2001）インターネット通販におけるカタログ再構成手法：洋服購入における評価プロセスモデル．電子情報通信学会技術研究報告，DE2001－39，p.17－24．
20）冨田一郎，手塚祐一，山本修一郎，長岡満夫（1998a）HTML文書からの商品情報抽出方式の提案．情報処理学会第56回（平成10年前期）全国大会講演論文集，**3**，p.79－80，2Y－3．
21）冨田一郎，手塚祐一，山本修一郎，長岡満夫（1998b）HTML文書からの商品情報抽出方式の提案．電子情報通信学会技術研究報告，KBSE97-27，p.15－22．
22）渡辺将尚，絹川博之，井田正明，芳鐘冬樹，野澤孝之，喜多一（2004）シラバスHTML文書からの情報抽出．情報処理学会第66回全国大会講演論文集，4，p.487－488．
23）山田信太郎，伊東栄典，廣川佐千男（2002a）Web上に公開されたシラバス情報の自動収集．情報処理学会マルチメディア，分散，協調とモバイル（DICOMO 2002）シンポジウム論文集，p.137－140．
24）山田信太郎，伊東栄典，廣川佐千男（2002b）自動収集したWebシラバスデータの分析と考察．情報科学技術フォーラム2002（FIT 2002）一般講演論文集第4分冊（N-32），p.301－302．
25）山田信太郎，松永吉広，伊東栄典，廣川佐千男（2002c）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），p.371－378．
26）山田泰寛，池田大輔，坂本比呂志，有村博紀（2004）WWWからの情報抽出：ウェブラッパーの自動構築．人工知能学会論文誌，**19**（3），p.302－310．

（受稿日　平成16年4月12日）

[ABSTRACT]

A Survey on Methods for Extracting Information from Web Documents: Considering their Application to Syllabus Data

YOSHIKANE Fuyuki*, IDA Masaaki**, NOZAWA Takayuki*,
MIYAZAKI Kazuteru*** and KITA Hajime****

In this paper, we summarize methods for extracting information from web documents. Some studies used methods based on linguistic information, while some studies used those based on structural information in HTML documents. As for syllabus data, it is desirable to adopt and combine both types of methods because of their linguistic and structural properties.

* Research Fellow, Faculty of University Evaluation and Research, National Institution for Academic Degrees and University Evaluation

** Associate Professor, Faculty of University Evaluation and Research, National Institution for Academic Degrees and University Evaluation

*** Associate Professor, Faculty for the Assessment and Research of Degrees, National Institution for Academic Degrees and University Evaluation

**** Professor, Academic Center for Computing and Media Studies, Kyoto University