成蹊大学理工学研究報告
J. Fac. Sci.Tech., Seikei Univ.
Vol.53 No.2 (2016)pp.25-28
（特別研究費に係る論文）

大学 Web ページからの研究室 Web ページの抽出

宮崎 敦也[*1]，酒井 浩之[*2]，坂地 泰紀[*3]

Extraction of Laboratory Web Pages from University Web Sites

Atsuya MIYAZAKI[*1], Hiroyuki SAKAI[*2], Hiroki SAKAJI[*3]

ABSTRACT：In this paper, we propose a method that extracts laboratory front pages from university web sites. Our method extracts the laboratory front pages by using SVM and applying some rules. Moreover, we developed the laboratory search system which is able to retrieve laboratory front pages extracted by our method. We evaluated our method and it attained 85.0% precision and 65.5% recall, respectively.

Keywords：text mining, web mining, information extraction

(Received October 21, 2016)

1．はじめに

平成 27 年度現在，779 もの大学が存在している[1]．大学への受験者は，自分の学力と大学の難易度を比べ志望校を決めている．しかし，大学で学べる内容も多岐にわたっており，自分の目指すことが出来る一番難易度の高い大学を受験し，そこに進学を決めた場合に，例えば，プログラムに興味がないにもかかわらず情報科へ進学するような進学ミスマッチもおこる．また，将来取り組みたい仕事を決めて，それについて勉強や研究をすることを目的に大学を選ぶ学生もいる．しかし，現在，大学ごとにどのような研究をしている研究室があるかをまとめた情報は存在しておらず，受検者が志望大学を選択する際に，自分の興味のある研究を行っている研究室が，どの大学にあるのかを探すことは困難である．

大学における研究室の情報を得るためには，大学の研究室Webページが有力な情報源となる．しかし，大学の研究室Webページへのリンクを大学Webページのトップページから探すことは手間がかかるうえに，大学ごとのWebページの構成が異なるため，すべての大学で同じ手順をたどって研究室Webページを閲覧できるわけでもない．また，大学の研究室Webページを「研究室一覧」という形で掲載している大学Webページもあるが，そのような大学はごく一部である．

そのため本研究では，大学Webページから研究室Webページを自動的に識別し，抽出することを目的とする．本研究では，まず大学Webページから，研究室Webサイトのトップページを抽出する（以降，研究室Webサイトのトップページを，研究室トップページと記述する．）．そして，抽出した研究室トップページからリンクをたどることで，その研究室Webサイトを構成するHTMLファイルを入手し，研究室Webページを抽出する．それにより，大学ごとに研究室Webページの一覧を自動的に生成できる．さらに，研究室Webページの HTMLファイル集合から抽出した情報を対象にした検索システムを構築した．それにより，大学受験者の興味のあるキーワードに関連する研究室を素早く検索することが可能となる．例えば「テキストマイニング」に興味があり，「テキストマイニング」の研究をしている研究室の一覧を得たい場合でも，本研究による検索システムにて「テキストマイニング」で検索することで，テキストマイニングを研究している研究室を検索できる．そのような情報は志望大学を決めるうえでの助けになると考える．

2．関連研究

関連研究として，酒井らは企業Webページから企業と関連のあるキーワードを抽出し，そのキーワードを検索

*1：情報科学科 学部学生
*2：情報科学科 准教授（h-sakai@st.seikei.ac.jp）
*3：情報科学科 助教

対象とした企業検索システムを提案した[2]．また酒井らは，企業の決算短信PDFから業績要因を抽出し，抽出した業績要因を検索対象とした決算短信検索システムを提案している[3]．文献[3]は，企業Webページから決算短信PDFをダウンロードできるIR情報ページを，SVMを使用して自動的に識別している．それに対し，本研究で対象とした大学Webページからの研究室Webページの抽出ではSVMのみでは困難であり，様々な規則を適用している．文献[2]に対して本研究は，大学Webページから研究室トップページのURLを自動的に抽出する必要があり，企業Webページ全体を使用した文献[2]の研究とは異なる．山田らは，シラバスデータの収集に着目し，効率良く収集する方法と収集したページの精度を向上する方法について考察している[4]．それに対して本研究では，大学Webデータに対してSVMを用いて研究室トップページを自動識別し，さらに，研究室トップページの特徴に基づいた規則を適用することで，研究室トップページを抽出した．

3．大学Webページからの研究室Webページの抽出

3．1　研究室トップページの抽出

本研究では，まず大学Webページから研究室トップページを抽出し，抽出した研究室トップページからリンクをたどることで研究室Webページを抽出する．本研究では737 大学分の大学Webページを取得し，ファイル数は合計で6,213,402 個のHTMLファイルを取得した．なお，1つの大学あたりの取得するHTMLファイル数の上限を25,000 とした．

大学Webページからの研究室トップページの抽出は，まずはSVMを用いて抽出を行う．ここで，学習用データとして，手動で収集した 36 個の研究室のトップページを選別し，素性として名詞を使用する．しかし，研究室のトップページのみでは素性として利用できる名詞が少ないため，手動で収集した研究室のトップページと，そのトップページに対するサブページ（トップページのリンク先のページ）を取得する．収集した36 の研究室のトップページに対して，そのサブページは合計 242 ページであった．そして，手動で選別した研究室トップページと，そのサブページに対して形態素解析を行い，名詞を抽出する．

次に，素性選択をするために，各名詞の出現頻度を求める．研究室トップページやサブページに多く出現している名詞は研究室トップページを識別する特徴的な名詞であると考える．しかし，出現頻度は高いが有用な素性にならないと判断される名詞（例えば，「日本」「田中」）もあった．そのような名詞はストップワードとして素性から除外した．そして，ストップワードを除く，出現頻度が 100 以上の名詞を素性とする．

手動で収集した 36 個の研究室トップページを正例とし，36 個の大学トップページを負例としてSVMで分類器を生成する．この際，1つの研究室トップページとそのサブページから1つの素性ベクトルを生成した．負例も正例と同様に，1つの大学トップページとそのサブページから1つの素性ベクトルを生成した．そして，737 大学の各大学Webサイトを構成するHTMLファイルをテストデータとし，SVMにより研究室トップページを抽出する．

3．2　URLリンクによる研究室トップページの選別

SVMにより抽出されたWebページには，研究室のトップページではなく，例えば研究室のメンバー紹介のような，研究室トップページの下位のページ（以降，研究室サブページと定義する）が多く含まれていた．そこで，精度の向上のために，SVMで抽出されたWebページに出現するURLリンクを取得し，多く出現するURLのWebページを研究室トップページとする．これは，研究室サブページに含まれるURLリンクには研究室トップページへのリンクが多く，その出現頻度が高ければ，研究室トップページである可能性が高いという仮定に基づく．以下に慶応大学のWebページから研究室トップページと識別されたページにおけるURLリンクとその頻度を取得し，頻度が上位のリンクのURLとその出現頻度を示す（表 1）．

表 1　取得したリンクの例

取得したリンクのURL	頻度
http://www.keio.ac.jp/	3692
http://www.keio.ac.jp/index-en.html	3611
http://www.keio.ac.jp/index-jp.html	2786

表 1 より，この処理において抽出されたURLは，実際には大学トップページや学部のトップページの出現頻度が高かった．そのため，この取得したURL集合とSVMにより抽出されたWebページのURL集合の積集合を，研究室トップページとする．これは，本手法によるSVMでは大学トップページが負例となっているため，生成された分類器では研究室トップページと識別されないからである．

3．3　規則による研究室トップページの選別

3.2 節で述べたように，SVMで抽出したWebページには，研究室トップページではなく，研究室サブページも多く含まれており．それらを除く必要がある．本節では，い

くつかの規則を適用することで研究室トップページの選別を行う．以下に選別手法を示す．

Step 1. 大学WebサイトからSVMにて研究室トップページと識別されたWebページのURLを / で区切り，要素数が少ない順にソート

Step 2. Step 1 にて並べられた順に，URLに対応するHTMLファイルのタイトルを取得

Step 3. Step 2 で取得されたタイトルに「研究室」,「Lab」,「LAB」,「lab」が含まれている場合，そのURLを研究室のトップページと判定．ただし，既に研究室のトップページであると判定されたURLを含むURLのページは，その研究室のサブページであると判定し削除．また，タイトルに「研究室紹介」,「研究室マップ」,「研究室一覧」を含む場合，そのページは研究室トップページでないと判定し削除．

Step 4. Step 2 からStep 3 を全てのHTMLファイルに対して繰り返す．

研究室トップページのURLの長さは一般的にそれほど長いものではない．そこで，Step 1 として，抽出されたWebページのURLからhttp://を除いた文字列を，/ を区切り文字として分割し，分割された要素の数が少ない順に並べる．そして，Step 3 として，研究室トップページであると判定されたURLを含むURLのWebページは，その研究室のサブページであると判定し削除した．例えば，表 2 において「http://iui.ci.seikei.ac.jp/」が研究室トップページとして判定されれば，以降，そのURLを含むサブページが抽出結果から削除される．

表 2　トップページとサブページの例

URL	ページ種類
http://iui.ci.seikei.ac.jp/	トップページ
http://iui.ci.seikei.ac.jp/internal/index.html	サブページ
http://iui.ci.seikei.ac.jp/member/index.html	サブページ

4．実装

本研究を実装し 737 大学のWebページから 9,115 の研究室トップページを抽出した．本手法の実装にあたり形態素解析器としてMeCab[1] を使用した．表 3 に成蹊大学のWebページから本手法にて抽出した研究室トップページのURLと研究室名をいくつか示す．

表 3　抽出された研究室トップページの例(成蹊大学)

研究室名	URL
知的インタフェース研究室	http://iui.ci.seikei.ac.jp/
材料力学研究室	http://www.sd.seikei.ac.jp/lab/zairiki/
言語情報研究室	http://www.ci.seikei.ac.jp/sakai/

さらに，抽出した研究室トップページを用いて研究室検索システムを作成した[2]．本システムは，研究室トップページと，トップページからリンクをたどった研究室サブページのHTMLファイル集合を研究室Webページとする．そして，そのHTMLファイル集合から文献[2]の手法を使用してキーワードを抽出し，そのキーワードを検索対象とした研究室検索システムである．図 1 に本システムにて「テキストマイニング」で検索した場合の検索結果を示す．

図 1 研究室検索システム

検索結果として，「テキストマイニング」と関連のある研究室と，その所属している大学が表示される．研究室名をクリックすると，その研究室トップページを閲覧することができる．なお，検索結果における研究室のランキングは，文献[2]のキーワード抽出手法によるキーワードのスコアに基づく．

5．評価

本研究の評価を行い，本手法の精度と再現率を求めた．

[1]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[2]http://hawk.ci.seikei.ac.jp/Lilas/

評価用の正解データは，ある大学の学科ごとに，その学科に属する研究室トップページをすべて人手で抽出して作成し，大学の学科ごとに本手法の精度，再現率を求めた．これは，１つの大学における研究室の数は多く，大学ごとに全ての研究室のトップページを人手で作成するには膨大な労力を要するからである．なお，正解データは７学科（７大学）から作成した．表４，表５，表６に，本手法Ⅰ，本手法Ⅱ，本手法Ⅲの精度・再現率の一部と７大学（７学科）の平均を示す（表４にのみ対象とした大学の学科名を併記する．）．ここで，本手法Ⅰは 3.2 節の処理を行わなかった手法，本手法Ⅱは 3.2 節，3.3 節の処理を行った手法，本手法ⅢはSVMによる抽出のみの手法（すなわち，3.2 節，3.3 節の処理をともに行わない）である．

表４ 本手法Ⅰの精度・再現率

	再現率	精度	F値
成蹊大学(情報科学科)	60.0%	85.7%	0.706
豊橋技術科学大学(情報・知能工学系)	81.4%	83.3%	0.824
東京工業大学(物理学科)	45.5%	76.9%	0.571
7 大学(7 学科)の平均	65.5%	85.0%	0.740

表５ 本手法Ⅱの精度・再現率

	再現率	精度	F値
成蹊大学	60.0%	100.0%	0.75
豊橋技術科学大学	46.5%	90.9%	0.615
東京工業大学	27.3%	85.7%	0.414
7 大学(7 学科)の平均	42.2%	90.9%	0.576

表６ 本手法Ⅲの精度・再現率

	再現率	精度	F値
成蹊大学	60.0%	2.84%	0.054
豊橋技術科学大学	86.0%	0.92%	0.018
東京工業大学	63.6%	3.07%	0.059
7 大学(7 学科)の平均	75.0%	6.47%	0.119

６．考察

評価結果から，本手法Ⅰによる再現率は約 65%，精度は約85%, F値は約0.740 と比較的, 良好な結果を示した．本手法Ⅱでは 3.2 節で説明した処理を行ったことにより精度は約 5%増加したが，再現率が 23%ほど減少してしまった．その理由として，3.2 節の処理により誤ったページが省かれたことにより精度が向上したが，研究室トップページも省いてしまったため，本手法Ⅰと比べると再現率が大幅に減少してしまったと考える．2 つの手法について，再現率と精度のF値を比較すると，本手法ⅠのF値約 0.740 に対して，本手法ⅡのF値は約 0.576 であったことから，本手法Ⅰの方が良好な結果であると言える．しかし，本手法Ⅰの再現率は約 65%であることから，抽出されていない研究室トップページも多くある．抽出されない原因は以下のように考える．

- 研究室トップページの HTML ファイルのタイトルに「研究室」や「Laboratory」が記述されていない．
- 研究室トップページであるが，ページが英語で記述されているため，本手法における SVM の素性が含まれておらず抽出されない．
- 大学 Web ページを取得する際に 1 つの大学あたりの取得する HTML ファイル数の上限を 25,000 とした．そのため，取得したデータに研究室のページがない．

以上の原因により, 研究室が抽出されていないと考える．そのため，まずは，より多くの大学Webページを取得すれば，再現率を上げることができると考える．

７．まとめ

本研究では，大学Webページから研究室Webページを抽出する手法を提案した．抽出にはSVMを用い，さらにいくつかの規則を適用した．評価の結果，本手法Ⅰでは再現率 65.5%, 精度 85.0%, 本手法Ⅱでは再現率 42.2%, 精度 90.9%を得ることができた．さらに，抽出した研究室トップページを使用して，研究室検索システムを構築した．今後は，より多くの大学Webページを取得し，そのデータに対して本手法を適用することで，再現率の向上を目指す．

参考文献

[1] 学校基本調査－平成 27 年度（確定値）結果の概要, http://www.mext.go.jp/b_menu/toukei/chousa01/kihon/kekka/k_detail/1365622.htm

[2] 酒井浩之, 坂地泰紀, “企業 Web ページを対象とした企業検索システムのための検索クエリに関連するタグの推定”, 第５回 テキストマイニング・シンポジウム, pp.41-45, 2014.

[3] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀, “企業の決算短信 PDF からの業績要因の抽出”, 人工知能学会論文誌, vol.30, no.1, pp.172-182, 2015.

[4] 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”, 電子情報通信学会論文誌 D , vol.J86-D1, No.8, pp.566-574, 2003.