様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ４ 月 １９ 日現在

機関番号：３２５０８
研究種目：挑戦的萌芽研究
研究期間：2014～2015
課題番号：２６５４０１８４

研究課題名（和文）ユーザ適応型オンライン教科書の自動生成に関する研究

研究課題名（英文）Study on the automatic generation of adaptive online textbook

研究代表者
柳沼　良知（Yaginuma, Yoshitomo）

放送大学・教養学部・教授

研究者番号：１０２５１４６４

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：　本研究の目的は，Web上に多量に散在する電子教材や学習オブジェクト等の教育コンテンツを自動収集，分析し，ユーザに適応した再構成を行うことで，ユーザ適応型オンライン教科書を自動生成する手法を開発することである．
　このため，まず，教育コンテンツのフィルタリングシステムを開発した．これは，Web上に存在する，教育コンテンツとして有用と考えられるWebページやPDFといったデータを多量に自動収集するものである．次に，これらのデータからの特徴抽出，分析，分類システムを開発し，教育コンテンツを意味内容の類似度に応じて関連づけて表示することを実現した．

研究成果の概要（英文）：The purpose of this study is to develop adaptive online textbook automatically. At first, a filtering system of educational content on the Web was developed. Then, browsing method which displays educational content according to the similarity was developed.

研究分野：情報工学

キーワード：オンライン教科書　プログラミングコンテンツ

１．研究開始当初の背景

OpenCourseWare や MOOCs に代表されるように，Web 上で教育コンテンツを公開する試みが広く行われるようになってきている．また，これら以外にも，学習のための素材として利用できるコンテンツが Web 上には数多く存在している．しかし，これらのコンテンツは，それぞれ個別に利用することを前提に作られ，散在しており，学習者が必ずしも容易にアクセスし，利用できる状況とは言えない．教育コンテンツの検索に，一般的な検索エンジンを利用することも考えられるが，検索結果には，通常，多くの無関係な情報が提示されるため，必要な情報に到達するまでに多くの時間や作業が要求されるという問題があった．

これらの教育コンテンツを利用するための基盤として，米国の MERLOT(Multimedia Educational Resources for Learning and Online Teaching)などの学習リソース共有・再利用機関が設立されている．これらは，検索に用いられるメタデータの登録や，ハーベスタソフトウェアによるメタデータの横断的な収集により，教育コンテンツの横断検索を実現している．しかしながら，それぞれのコンテンツ内の部分部分の内容に応じた検索や，複数のコンテンツを関連づけて利用するといったことは想定されておらず，学習者が必ずしも容易にアクセスし，利用できる状況とは言えない状況があった．

２．研究の目的

本研究では，Web 上に多量に散在する電子教材や学習オブジェクト等の教育コンテンツを自動収集,分析し、ユーザに適応した再構成を行うことで,ユーザ適応型オンライン教科書を自動生成する手法を開発することを目指す．

具体的には，まず，教育コンテンツのフィルタリングシステムを開発する．これは，Web 上に存在する，教育コンテンツとして有用と考えられる Web ページや，Word，Excel，PDF といったデータを多量に自動収集するものである．また、教育コンテンツは，文字情報だけでなく，２次元画像，動画像等のマルチメディア情報が利用されるようになってきていることから，こういった教育コンテンツとして利用可能なマルチメディアデータの収集もあわせて行う。

次に，これらのデータからの特徴抽出，分析，分類システムを開発し，教育コンテンツを意味内容の類似度に応じて関連づけて表示できるようにするとともに、それぞれの学習者に適応した表示方法を実現する．これにより，本研究では，Web 上の教育資源を有効に活用し，適応的な学習環境を実現することを目指す．

３．研究の方法

平成 26 年度は，

(1)理論的枠組みとツールの先行研究調査

(2)Web 上の教育コンテンツの自動収集

を行い，提案システムの設計を行った．

具体的には，(1)理論的枠組みとツールの先行研究調査については，OpenCourseWare や MOOCs といった Web 上の教育コンテンツの配信の形態や，米国の MERLOT 等の学習リソース共有・再利用機関に関する文献調査や Web 調査を行うことで，処理対象となるデータの特性を明らかにした．また，近年注目されているビッグデータの解析手法の適用事例の収集等により，ユーザ適応型オンライン教科書の自動生成システムの設計を行った．

また，(2)Web 上の教育コンテンツの自動収集に関しては，近年，大学等で，電子教材やそれに付随する関連データを公開する動きが広がっているが，通常，それらは，Web 上に独立に散在している．このため，それらを関連づけ，ユーザに応じた利用を実現するためには，まず，ある Web ページが「教育コンテンツに関するページ」であるか否かを判別する仕組み（フィルタ）を開発することが必要となる．その予備段階として，HTML だけではなく Word，Excel，PDF，画像，動画など様々なフォーマットの教育コンテンツを収集し，教材構造やよく用いられるキーワードや特徴量などの分析を行った．

平成 27 年度は，

(3)収集した教育コンテンツの分析・体系化手法の開発を行った．

教育コンテンツを高い精度で多量に収集できたとしても，そのままでは，必ずしも学習者にとって必要な情報を，理解しやすい形で提示できるわけではない．このため，収集した教育コンテンツをその学習内容に応じて，自動的に関連づけ，分類・提示する手法を開発する．分類・提示手法を確立することで，教育コンテンツを体系化でき，学習者は，自分が学びたい内容を容易に選び出し，利用することができるようになる．

教育コンテンツの分析，分類には，テキスト内で用いられているキーワードの出現頻度を用い，学習者にとって利用しやすい，検索，閲覧インタフェースの検討をあわせて行った．

４．研究成果

(1) プログラミングコンテンツの収集

Web 上にあるプログラミングコンテンツの収集について検討するために，検索エンジンで検索した結果に，どのようなコンテンツが含まれるかの分析を行った[1,2]．

手順としては，まず，検索エンジンで，「プログラミング」のキーワードで検索を行い，検索結果のリストのうち，上位 300 を抽出した．次に，これらの URL のリンク先の HTML ファイルのダウンロードを行うとともに、文字コードを揃えるために，シフト JIS に変換する処理を行った．(276 のファイルについて，エラーが出ずにこれらの処理を行うことが

できた.)

次に、ダウンロードした HTML ファイル中のタグや，Web ページのデザインのためのスタイルや，処理を行うための JavaScript の記述は，テキスト処理を行う際に不要な情報であることから，Java のプログラムを作成し，削除する処理を行った．

このように収集した Web ページに対して，文章を単語ごとに分割する形態素解析の処理を行い，出現頻度が高い順に，約 200 の名詞を選択した．そして，ページごとに，これらのキーワードの出現頻度のベクトルを作成した．この出現頻度のベクトルを特徴量として，対応分析を行い，2 次元平面上に各ページの配置を行った．

図 1 は，「プログラミング」で検索された Web ページに対して，対応分析を行った結果である．右下にページが集中する部分があり，この部分には，″Java″，″PHP″といったプログラミング言語名や，″入門″，″講座″といった名詞が並んでおり，プログラミングや講座等の説明がなされている Web ページが集まっている部分と考えられる．左上には「夫」のキーワードがあるが，これは，タイトルに「夫」を含む Web ページへのリンクを多量に含むページがあったことで，それが反映されたものである．左下には，″/*″や″*/″，″{″や″}″，″//″など，ソースコードの中に多く含まれる文字列が配置されていることから，この部分には，主に，具体的なソースコードについて述べている Web ページが存在していると考えられる．

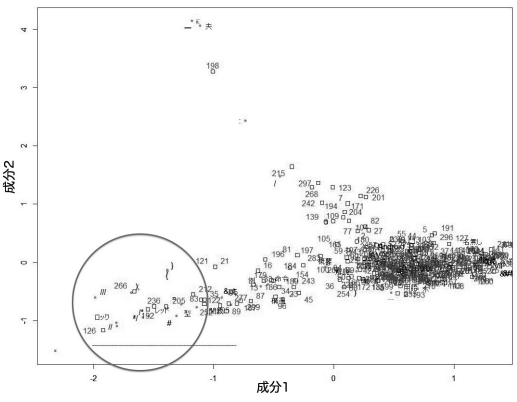

図 1　「プログラミング」の分類結果

同様に，Web 上にあるプログラミングコンテンツの収集を実現するために，検索エンジンで検索した結果に，どのようなコンテンツが含まれるかの分析を行った．具体的には，検索エンジンで，「C 言語 プログラミング」のキーワードで検索を行い，検索結果の 177 の Web ページを HTML 形式で保存した．そして，ダウンロードした HTML ファイルの文字コードをシフト JIS に変換し，また，タグやスタイル，JavaScript の削除を行った．このようにして抽出したテキストデータに対して，形態素解析ソフト ChaSen を用いて，文書を単語ごとに分割し，出現頻度が高い順に，約 200 の名詞を選択した．そして，これらのキーワードの出現頻度のベクトルを特徴量として，対応分析を行った（図 2）．

この結果を見ると，左上の，″/*″や″*/″，″//″といったソースコードに特有なキーワードの近くには，実際に何らかの C プログラムのサンプルが載っているなど，C 言語プログラミングに関連性の高い Web ページが配置されることが分かった．

次に，Web 上からまとまった内容を網羅的に収集することを目的に，前述の Web ページの中から 17 のページを選び，それらのページを種として，再帰的に 2 段階のリンクをたどり，Web ページの収集を行った．2 段階のリンクをたどることで，あるページがプログラミングに関するページの一部だった場合，そこから目次のページをたどり，そこから更にリンクされているプログラミングコンテンツをたどるといったことが可能になる．以上の処理により，約 5400 のファイル，約 190MB のデータが得られた．この中で，HTML ファイルは，約 3100 あった．この HTML ファイルを対象に，その表示手法について検討する．

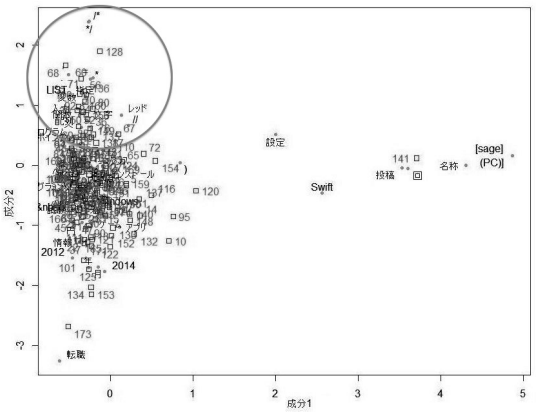

図 2　「C 言語 プログラミング」の分類結果

(2)プログラミングコンテンツの表示手法

収集した Web ページの表示は，3 次元空間の中で，類似するページ同士が近くになるように配置することで行った[3]．これにより，それぞれのページの関連性を視覚化することができる．

それぞれの Web ページの 3 次元空間中での座標については，対応分析により求めた．具体的には，まず，文字コードを揃えるために，それぞれの Web ページの文字コードをシフト JIS に変換する処理を行った．また，ダウンロードした HTML ファイルの中には，タグや，Web ページのデザインのためのスタイルや，処理を行うための JavaScript の記述があることから，Java プログラムを作成し，これらを削除する処理を行った．具体的には，″<style″から″</style>″，″<script″から″</script>″の間の文字列を削除することで，スタイルと JavaScript の削除を行った．また，タグについては，<(″[^″]*″|'[^']*'|[^'″>])*>の正規表現を

用いて，"<"と">"の間の文字列を削除した．この結果，約 65 万行，24MB のテキストデータを得ることができた．

このようにして，テキスト部分のみを抽出した後,形態素解析ソフト ChaSen を用いて，単語ごとに分割する処理を行い，出現頻度が高い順に,約 200 の単語を選択した.そして，ページごとに，これらのキーワードの出現頻度のベクトルを作成し，この出現頻度のベクトルを特徴量として，対応分析を行った．対応分析は，テキストマイニングなどで広く利用される手法であり，キーワードの頻度の和を 1 になるように正規化し，その類似度が高いものが近くになるように配置する手法である．この時，関連するキーワードを重ねて表示することができ，それぞれのページがどのキーワードに関連が深いかを知ることができる．ここでは，対応分析結果の上位 3 軸を選び，その 3 次元座標を以下の処理に利用した．

対応分析結果の表示には，X3D を利用した．X3D は,3 次元空間を記述する言語である VRML の後継となる言語であり，XML ベースで 3 次元空間の記述を行うことができる．X3D ファイルの表示には，X3DOM を利用した．X3DOM は，X3D を WebGL をベースに表示するための JavaScript ライブラリであり，Web ブラウザ上で,X3D データの表示を行うことができる．

出現頻度が高い単語を表示する場合，問題となるのは，文字は平面上に表示されるため，全体を水平方向に 90 度回転させた場合，幅が 0 となり，単語を見ることができなくなる点である．X3D では，各オブジェクトに対して，各軸を中心に回転を許すか許さないかを指定することができるため，ここでは，オブジェクトの回転を許さないという設定にすることにより，全体を回転等しても各オブジェクトは常に正面を向くようにした．

各 Web ページの表示については，それぞれのサイトごとに，色（赤，黄，緑，シアン，青，マゼンタ等）を変えることで，それぞれのページがどのサイトに含まれるものかを見やすくしている．また，<anchor>タグにより各オブジェクトのリンク先を指定し，それぞれのオブジェクトをクリックした場合に，対応した Web ページを開けるようにした．

作成した X3D ファイルを表示させると図 3 のようになる．ビューアの機能により，ユーザは，この 3 次元空間に対して回転，拡大縮小，平行移動等の操作を行うことができる（図 4）．そして，それぞれのオブジェクトをクリックすることで，対応する Web ページを表示することができる．

今後は，より多くのデータの収集を行い有効性を検証するとともに，より効率の良い表示方法や，より意味内容を反映させた表示方法等について検討を行っていく予定である．

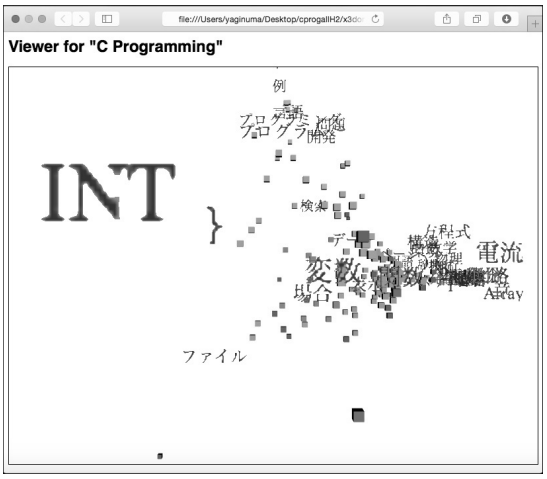

図 3　プログラミングコンテンツの表示

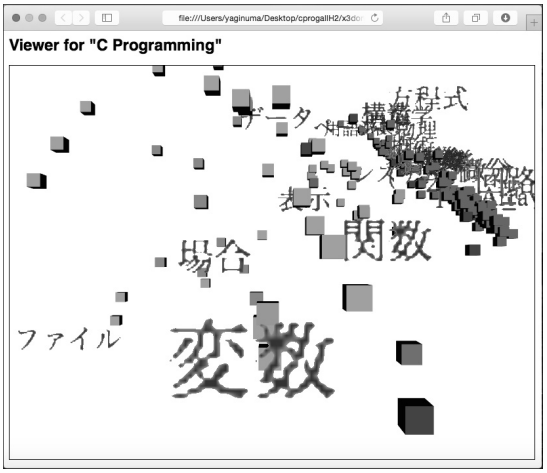

図 4　表示画面の操作（拡大）

５．主な発表論文等

〔学会発表〕（計　3　件）

[1]柳沼良知："Web 上のプログラミングコンテンツの収集と分類"，電子情報通信学会技術研究報告，114 (260)，pp.31-34 (2014.10.18) 金沢大学

[2]柳沼良知："Web 上のプログラミングコンテンツの収集手法の検討"，教育システム情報学会研究報告，30(1)，pp.89-92 (2015.5.23) 放送大学

[3]柳沼良知："Web 上のプログラミングコンテンツの表示手法の検討"，電子情報通信学会技術研究報告，115(492)，pp.59-60 (2016.3.5) 香川大学

６．研究組織

(1)研究代表者

柳沼良知（YAGINUMA, Yoshitomo）

放送大学教養学部・教授

研究者番号：10251464