様式 C-19

科学研究費補助金研究成果報告書

平成２２年５月３１日現在

研究種目：若手研究（B）
研究期間：2007～2009
課題番号：19700657
研究課題名（和文）　メタデータ付与の自動化と講義情報の配信による教育用コンテンツの流通促進

研究課題名（英文）　Distribution of Educational Contents Based on Offering Course Information by Using Automated Metadata Generation

研究代表者
辻　靖彦（Yasuhiko TSUJI）
放送大学・ICT 活用・遠隔教育センター・准教授
研究者番号：10392292

研究成果の概要（和文）：

本研究は、教育用コンテンツ、特に講義シラバスを対象にメタデータ（LOM）を自動生成すること及び、LOM を用いた講義情報を配信するシステムを開発することで教育用コンテンツの共有を促進させることを目的としている。LOM を自動先生するために、講義シラバスページから情報抽出を行う手法、講義情報の専門分野への自動分類手法、最新の講義情報一覧ページからスクレイピングを用いて情報抽出を行う手法を開発した。各手法について精度を確認した所、それぞれ 93.8%、87.7%の精度が確認された。さらに、最新の講義情報一覧ページから Web スクレイピングを用いて情報抽出を行う手法及び、LOM を用いた講義情報配信するためのプロトタイプシステムを開発した。

研究成果の概要（英文）：

The purpose of this research is to distribute of educational contents based on offering course information by using automated metadata generation. We developed an information extraction method of course syllabi to make LOM items automatically, and an automatic syllabus classifier to make “classification” item of LOM. And we analyzed the accuracy of each method. As a result, we could find the accuracy rates 93.8% and 87.7%. In addition, we proposed an information extraction method from syllabus list pages on the web by using web scraping and developed a prototype system of offering course syllabi using LOM.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	900,000	0	900,00
2008 年度	500,000	150,000	650,000
2009 年度	1,700,000	510,000	2,210,000
年度			
年度			
総　計	3,100,000	660,000	3,760,000

研究分野：教育工学
科研費の分科・細目：科学教育・教育工学―教育工学
キーワード：シラバス、情報抽出、メタデータ、LOM、文書分類

１．研究開始当初の背景

インターネットの普及と共に、様々な専門分野の教育用コンテンツが大学、学校、法人、企業、あるいは個人によって開発され、インターネット上で公開されている。そして今後さらに増大し続けると考えられる。しかし、教育用コンテンツは各組織のサーバに分散して置かれていることから、各コンテンツに対して“タイトル”、“概要”、“専門分野”などを記した学習対象メタデータ（以下 LOM と記す）を付与することは大変重要である。なぜなら LOM を用いることにより、分散している各コンテンツを横断的かつ体系的に検索することが可能となるからである。例えば、「ディジタル信号処理」というキーワードを入力することでそのキーワードを含む全国の教育用コンテンツを一度に検索することが可能となり、“専門分野”が「情報工学」分野の教育用コンテンツに対してキーワード「遺伝的アルゴリズム」を“概要”の中に含むもののみを検索することも可能となる。これらの LOM を用いた検索機能は独立行政法人メディア教育開発センター（現在の放送大学 ICT 活用・遠隔教育センター）で開発した能力向上学習ゲートウェイ NIME-glad（http://nimg-glad.code.ouj.ac.jp/）で既に実装されており、131,629 件の教育用コンテンツが検索可能である。

しかし、講義シラバスなど Web 上の教育用コンテンツに対して LOM を付与することは、知識を持った人が手作業で行わなければならないため大きな作業コストがかかる。そのために支援システムの開発及び自動化が求められているが、これまでに実用的なシステムは実現されていない。特に講義シラバスは少なくとも年に１度は更新される特徴を有するため、LOM の更新が自動化されていることが望ましいと考えられる。

また現状では、講義シラバスをデータベース化して一元的に管理している大学も多い。しかし、シラバス登録システムを導入・運用するには大きなコストがかかるのが現状である。また、登録システムを持たない大学では教員自身が自分の研究室等の Web ページで講義シラバスを作成しなければならず、その場合は教員に大きな負担がかかってしまう。

２．研究の目的

上記の背景の下に、本研究で達成する目標は以下に示す２点である。

（１）全国の大学のシラバスの Web ページから LOM の項目情報を抽出する手法の開発と評価

本研究の目的は、インターネット上の講義用のシラバスを対象に LOM を自動的に生成することである。そのためには、講義シラバスから“タイトル”、“概要”、“専門分野”などの LOM 項目情報を取得する必要がある。初めに LOM 項目情報の自動抽出エージェントを開発する。そして全国の大学等のシラバスを対象に、大規模な抽出の正確性について評価を行う。また、“専門分野”の情報については、情報抽出ではなく、シラバスの文書中に出現する専門用語から分野を判定する文書分類の手法が必要となる。従って、講義シラバスの自動分類手法についても開発する必要がある。

（２）シラバス登録・配信システムの開発

LOM 付与の作業負担の軽減と講義シラバスの作成・LOM 登録・公開サービスを目的とした授業シラバス配信システムを開発する。本システムを用いて講義の内容を Web ブラウザ上から一括登録することにより、IT スキルを持たないユーザでもシラバスを Web 上へ登録し、公開することが可能となる。作成したシラバスには URL が与えられ、インターネットで自由に閲覧できる。また、シラバスのレイアウトやデザインも修正できる。さらに、ユーザが入力する項目は LOM に対応している項目を用いるため、シラバスの作成と同時に LOM も生成でき、さらに検索システムへ登録できる点が特徴である。また、本システムは、PC 及び携帯電話の双方で利用できる形で開発を行う。

以上、本研究で開発する LOM 自動生成システムを用いる事で、全国の大学の講義シラバスに対して人手を介さずに LOM を生成し、検索システムに登録することができるようになるため、教育用コンテンツの流通・共有化が大幅に促進されると期待できる。また、Web ベースのシラバス配信システムを用いることで、大学等の講義シラバスを Web ページ化して公開する技術的作業及び LOM 付与の作業負担が無くなり、大学の教職員が直接シラバスを登録できる。従って、講義情報の流通・共有化が大幅に促進されると期待できる。

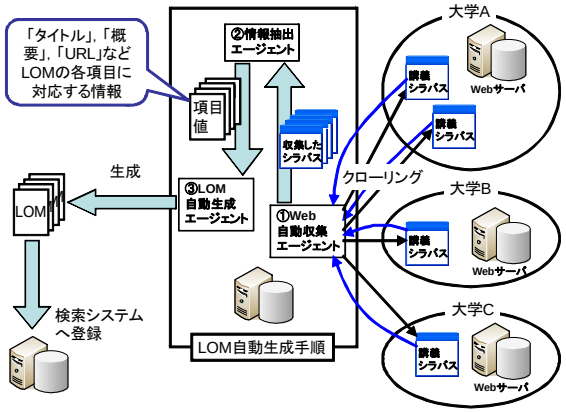

図 1. LOM 自動生成の流れ

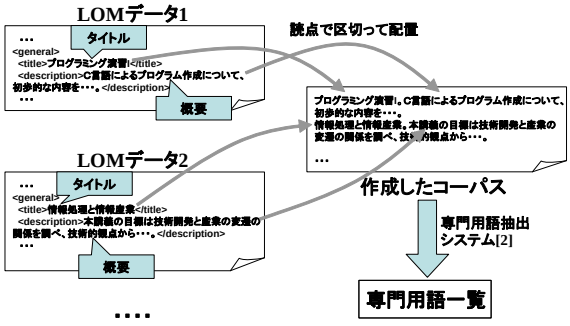

図 2. 専門用語の抽出

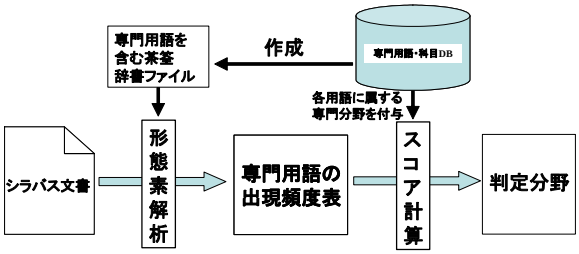

図 3 自動分類の流れ

3．研究の方法

本研究の方法を以下に説明する。

3．1 シラバスの Web ページから LOM の項目情報を抽出する手法の開発

本研究で用いる LOM 生成手法の全体構成を図 1 に示す。図 1 より、本研究では「①Web 自動収集エージェント」によりインターネット上のシラバスを自動的に収集し、「②情報抽出エージェント」で収集した各シラバスから LOM に必要な項目情報を抽出し、「③LOM 自動生成エージェント」で項目情報を基に所定の XML 形式の LOM フォーマットに変換を行うことで LOM の自動生成が実現可能であると考える。①については Web クローリング技術及び、「授業科目名」などのシラバス中で用いられる用語の頻度に着目する既存の手法を用いて実現可能と考えられる。また、③については項目情報を特定の XML 形式に変換すればよいので実現可能である。②については、「タイトル」、「概要」、「URL」など LOM の各項目に対応する情報を、様々な大学、学科のなるべく多くのシラバス文書から高精度で自動的に抽出できる事が求められる。そこで本研究では、 LOM の項目の「タイトル」として利用可能な「授業科目名」属性及び、LOM の項目の「概要」として利用可能な「授業目的・内容」属性の両者について自動的に情報抽出を可能とする手法を開発することを目的とし、手法の開発を行った。初めに単純な抽出アルゴリズムを用いて情報抽出の予備実験を行い、精度を確認した。続いて抽出に失敗したシラバスの書式を加味した上でアルゴリズムの改良を行い、抽出精度を確認した。

3．2 シラバス文書からの自動分類手法の開発

LOM 項目の一つ“専門分野”を自動生成するために、シラバス文書中の専門用語の出現頻度を用いた専門分野への自動分類手法の開発を行った。その流れを以下に示す。

本研究で開発した自動分類手法では、講義シラバスの内容から専門分野を判定するために、専門用語に着目している。そこで、NIME-glad へ登録されている、21 分野 81,286 件の講義シラバスの LOM データを用いて専門用語を抽出した。専門用語を抽出する流れを図 2 に示す。初めに各分野の LOM データから、タイトルと概要を抜き出して「＜タイトル＞。＜概要＞。」の形に変換してコーパスとみなす文書を作成した。そして茶筌で形態素解析を行い、名詞の連結頻度に基づく抽出手法を用いて 21 分野の専門用語を抽出した。ここで、専門分野は文学、教育学、社会学、法・政治学、経済学、医・歯・薬・保健看護、農学・獣医学、家政学、芸術・音楽、体育学、電気電子、情報、建築・土木、機械、材料、システム・制御、生物化学・バイオ、数学、物理、化学、地球・環境科学である。

抽出した専門用語を基に開発した自動分類手法を図 3 に示す。新しいシラバス文書を入力すると、専門用語 DB から作成した専門用語を含む茶筌辞書ファイルを用いて形態素解析を行う。そしてシラバス文書の中にどの分野の専門用語が出現するかをスコア付けし、21 分野の中で最もスコアが高い分野を判定結果とする。専門用語 DB 及び辞書ファイルは、LOM データから抽出した専門用語の上位 700 語×21 分野の合計 14,700 語を用いた。スコア付けの計算としては、①出現頻度そのものの合計、②（出現頻度／出現分野数）の合計、③複数分野に出現する用語は除いた場合の用語の出現頻度の合計、の 3 通りの計

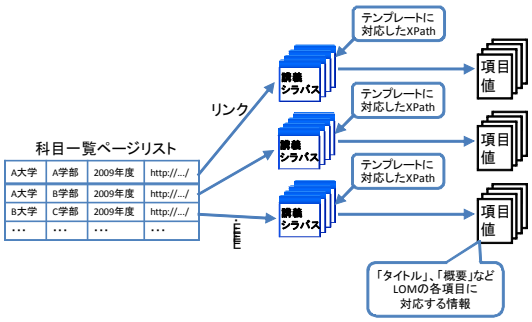

図 4. 最新シラバス一覧ページからの情報抽出手法の流れ

算方法について検討、比較した。

続いて、作成した辞書に対して、用語の精査を行った。21 分野から抽出した上位 1000 語の用語を対象に、初めに前処理として「情報基礎演習Ⅲ」の『Ⅲ』や「社会心理学特講 A」の『A』など授業科目名に付与している記号を削除し、続いて用語自体のチェックを行い、専門用語として適切では無いものを削除した。削除した理由は以下の 8 通りであった。

①授業で一般的
②研究で一般的
③科学技術で一般的
④実験用語
⑤他分野と思われる用語
⑥接頭語・接尾語
⑦1 文字の用語
⑧その他汎用語

各分野の上位 1000 語から①～⑧に含まれる用語を削除した後、残った用語から上位 700 語、21 分野で合計 14,700 語を精査後の新しい辞書とした。尚、削除用語は全ての分野から削除を行っている。例えば「講義」という用語を削除すると決定した際には、21 の全ての分野の用語辞書から「講義」の語を削除した。また、⑥に含まれる用語については、重複しない限り、「～等」や「～系」などの接頭語や接尾語のみ削除した。

3．3 全国の国公立大学の最新シラバス一覧ページからの情報抽出手法の開発

大学の最新のシラバスの一覧ページから情報抽出を行い、XPath 及び Web スクレイピングにより情報抽出を行う手法を開発した。提案する手法を以下に示す（図 4 参照）。

Step 1. 全国の国公立大学の大学・学部・研究科の Web サイトから、最新の各講義のシラバスページのリンクを有する「授業科目一覧ページ」を手動で取得し、科目一覧ページリストを作成
Step 2. Step 1. で取得した各ページから各講義シラバスの URL を自動取得
Step 3. Step 2. で得た講義シラバス群に対して、必要な情報の位置を示す XPath を手動で作成
Step 4. Step 3. で得たシラバス文書と XPath に基づいて情報抽出

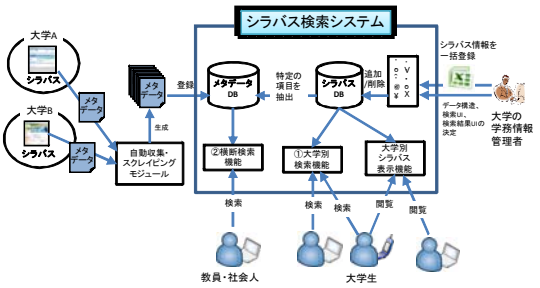

図 5. シラバス登録・配信システム

ここで、Step1. で作成する科目一覧ページリストは、HTML 形式のシラバスへのリンクを有するページのみを対象とし、PDF などの他のフォーマットの場合や、検索システムとなっている場合は除外する。また、Step 3. で作成する XPath は、「抽出したい項目名＋XPath」の形式を取り、同じ一覧ページからリンクされているシラバス群に対しては同じ XPath を適用する。その理由は、同じ科目一覧ページからリンクされているシラバスは、似たような HTML フォーマットを持つと考えられるからである。

3．4 LOM 検索システムとの連携を考慮したシラバス登録・配信機能を有するプロトタイプシステムの開発

本研究で提案するプロトタイプシステムの構成を図 5 に示す。本システムは①大学別シラバス検索機能及び②シラバスの横断検索機能の 2 つの機能を有している。①はその大学（又は学部・学科）ごとにシラバスを検索できる機能であり、検索インターフェイス、検索結果インターフェイス、シラバス表示インターフェイスを個別に設定できる点が特徴である。これらの設定及びシラバス情報のアップロードは大学等の学務情報管理者が行う事を想定している。シラバスが登録された段階で②の横断検索用のメタデータも並行して生成され、メタデータの DB にも登録される。大学別検索機能の利用者は各大学の教職員及び学生を想定している。学生の利用を推進させるために、PC だけでなく携帯電話でも利用可能とする。②は本システムと Web 上のシラバスの両方を横断的に検索することができる機能である。Web 上のシラバスページから作成したメタデータ及び、本システムに登録された各大学のシラバスページから作成されたメタデータの双方を検索できる。利用ユーザとしては主に、大学教職員及び社会人を想定している。尚、②の機能につ

いては検索は本システムで行うが、閲覧する際には大学のサーバへアクセスするので、携帯電話への対応はしていない。

４．研究成果

４．１ シラバスの Web ページから LOM の項目情報を抽出する手法の評価

開発したアルゴリズムを用いて、ac.jp の 84 ドメインの 12066 件のシラバスを対象に抽出精度を確認した。その結果、「授業科目名」属性については 10513 件、「授業目的・内容」属性は 7159 件の抽出結果が得られた。そのすべての結果事例について、目視により精度を確認した。「授業科目名」属性の抽出に関しては、10513 件の抽出結果に対して、9863 件の授業科目名が正確に抽出できた。抽出精度は 93.8%（9863/10513）であった。正確に抽出できた内訳は、項目名が「科目名」であるのが 9212 件、項目名が「授業科目」であるのが 651 件であった。抽出に失敗した 650 件に対して原因を目視で確認した所、シラバスが Microsoft の Word により作成された HTML 形式であるのが半数以上の 343 件を占めていた。Word で HTML ファイルを作成すると HTML:タグが複雑な構造となるために抽出が困難であった原因が考えられる。他の失敗例としては、本抽出で想定していない「科目名副題」や「本授業科目に関する情報」など、授業科目名の項目名以外の文章にヒットしてしまう事例が 268 件確認できた。尚、HTML のテーブル構造の縦・横を誤判定する事例は 20 件だけであった。授業目的・内容属性については、7159 件に抽出結果に対し、6281 件の授業目的・内容が抽出できた。抽出精度は 87.7%（6281/7159）であった。抽出成功した項目名は「達成目標」が 3261 件で最も多く、それに続いて「講義概要」1013 件、「講義目的」617 件、「授業の目的」404 件、「講義内容」244 件となった。抽出が失敗した例としては、授業計画の表の中に「授業内容」などの項目名にヒットして誤抽出してしまう事例が最も多く、317 件あった。他の例としては、<p>タグ、<font>タグ、<dd>タグなどで実際の授業目的が区切られているために一部のみしか抽出できない事例が多く見られた(262 件)。他には授業科目名属性と同様に、Word で作成されたシラバスであるために抽出に失敗している事例が 189 件確認された。

４．２ シラバス文書からの自動分類手法の評価

3.2 節で開発した自動分類手法について、評価を行った。3 大学 4 学科（情報工学科、電気電子工学科、法律学科、経済学科）に属する合計 357 件の講義シラバスに対して自動分類を試み、分類が正しいか確認した。その結果、出現頻度による①の方法では 82.1%（357 中、293 シラバスが分類に成功)、②の方法では 89.1%（318/357)、③の方法では 88.8%（317/357）の精度で自動分類に成功した。これより、②のスコア計算方法が最も優れている結果となった。

表 1 自動分類の実験結果

判定結果 / 分野	情報	電気電子	法・政治学	経済学	その他
情報	85.3% 64/75	6/75	0	0	5/75
電気電子	0	84.8% 39/46	0	0	7/46
法・政治学	0	0	92.3% 72/78	0	6/78
経済学	1/78	0	3/78	85.9% 67/78	7/78
その他	0	0	0	0	95.0% 76/80 ※1

数値はシラバスの件数
分野と判定結果が同じセルは
精度(%)
正解シラバス数／全シラバス数

※1 4件が判定失敗

また、②の方法における、分野別の分類結果を表 1 に示す。この表より、『情報』と分類されるべきシラバス 75 件のうち、64 件が正しく「情報」に分類され、6 件が「電気電子」、5 件が 4 分野以外の「その他」の分野へ誤って分類されていた。「その他」の分野とは、「システム制御」・「家政学」・「美術音楽」であった。他の誤答としては、『電気電子』のシラバスは 46 件中 7 件において「物理」・「材料」・「システム制御」へ誤答、『法・政治学』は 78 件中 6 件が「文学」への誤答、『経済学』のシラバスは 78 件中 1 件が「情報」、3 件が「法・政治学」、7 件が「文学」・「数学」・「建築・土木」・「農学・獣医学」への誤答が確認された。この理由としては、その分野の専門性と関連の低い用語が多く含まれている可能性があるためである。

続いて、辞書に含まれている用語の精査を行い、精査後の用語辞書を用いてシラバスの分類実験を行った。確認用文書として、ac.jp ドメインから 2008 年 6 月にクローリングした中からランダムに選択した、21 分野の 1964 件のシラバス文書を用いた。こちらを用いて、既に開発している精査前の辞書との精度の比較も同時に行った。

分類する手順は、１．専門用語辞書から形態素解析用の茶筌辞書ファイルを作成、２．シラバス文書を形態素解析、３．１で作成した辞書ファイルから専門用語の形態素を抽出、４．各分野に対して、属する専門用語の出現頻度に応じてスコアを集計、５．最もスコアの大きな分野を判定結果として出力する。ここで、スコアの計算式には【(TF／属

する分野数）の合計値】を用いた。また、自動分類した結果が正解かどうかの判定は実験者が行った。シラバスの内容によっては複数の分野に属するものも考えられるが、その場合は、自動分類の結果がその中のいずれか 1 つを出力できれば正解とした。結果としては、精査前の辞書を用いた自動分類では精度が 83.6%（1641/1964）であったのに対し、精査後の辞書では 87.7%（1722/1964）の精度が確認でき、精度が 4.1%改善された。

４．３全国の国公立大学の最新シラバス一覧ページからの情報抽出手法の評価

全国の国公立大学 42 大学から 1361 のシラバス一覧ページを抽出し、そこから本手法を用いて情報抽出を試みた。その結果、65,535 件の講義シラバスのメタデータを抽出できた。抽出精度の評価については今後の課題である。

４．４ LOM 検索システムとの連携を考慮したシラバス登録・配信機能を有するプロトタイプシステムの評価

本システムはプロトタイプシステムとして開発したため、システムの評価及び、実際の公開・運用については今後の課題である。

５．主な発表論文等

〔学会発表〕（計 5 件）

辻 靖彦、清水康敬, LOMデータから作成した専門用語辞書の精査と講義シラバスの自動分類, 電子情報通信学会（第二種研究会資料, SIG-WI2-2008-70, pp.85-86), 横浜・神奈川近代文学館, 2008.12.12

辻 靖彦, 森本容介, LOMの自動生成を目的としたシラバス文書の情報抽出, 電子情報通信学会教育工学研究会（信学技報, vol. 109, no. 335, ET2009-74, pp. 131-136), 琉球大学, 2009.12.11

辻 靖彦, 森本容介, メタデータの自動生成を目的とした最新シラバスからの情報抽出, 教育システム情報学会（研究報告, Vol.24, No.6, pp.142-143), 畿央大学, 2010.3.13

森本容介, 辻 靖彦, 山田恒夫, 学習コンテンツのメタデータ検索エンジンの開発, 電子情報通信学会（技術研究報告, Vol.110, No.42, pp.13-17), 琉球大学, 2009.12.11

辻 靖彦、清水康敬, LOMデータから抽出した専門用語に基づく講義シラバスの自動分類, 日本教育工学会第 23 回全国大会（講演論文集 1p-115-05 , pp.343-344), 早稲田大学所沢キャンパス, 2007.9.23

〔図書〕（計１件）

辻 靖彦、亀井 美穂子, 学びとコンピュータハンドブック―2.2 節　情報検索-（分担執筆）

６．研究組織

(1)研究代表者

辻　靖彦（Yasuhiko TSUJI）

放送大学・ICT 活用・遠隔教育センター・准教授

研究者番号：10392292