様式 C-19

科学研究費補助金研究成果報告書

平成　24 年　6 月　30 日現在

機関番号：32692
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：20500848
研究課題名（和文）既存教材よりのオントロジー構築と定量化

研究課題名（英文）An Ontorology Construction and its Measure from Teaching Materials

研究代表者
塚本　享治（TSUKAMOTO MICHIHARU）
東京工科大学・メディア学部・教授
研究者番号：90386764

研究成果の概要（和文）：　大学の学部で公開されているカリキュラム、シラバス、講義・演習資料、およびインターネットから入手できる資料を取り込んでそれらに現れる重要用語の関係を RDF 化するソフトウェアを開発し、大量の教材を収集し RDF 化した。そののち、科目、教材、用語などの関係を OWL 言語で記述して、RDF に現れない関係を導出できるオントロジーを構築した。このオントロジーを Sparql 文で検索するタグライブラリを備えた汎用の Web アプリケーションサーバを実現した。このサーバシステムを使って教材群に対して科目間の関係の定量化を試みた。オントロジーからは思うような論理的結果がえられなかったが、用語の出現頻度に着目する TF-IDF 法をつかって教材間に潜む一般的傾向を導出した。

研究成果の概要（英文）：　Our faculty has provided the curriculum, syllabuses, teaching materials and a timetable on its home page. We developed the software to gather them, analyze in morpheme using Japanese analyzer “KoBaKo”, and convert to RDFs. Then we described relationships among terms in OWL language. The system searched original RDFs and deviated RDFs by Sparql queries. We have developed the server system “meiseki” which provide such information processing services. Using the integrated system, we gathered teaching materials and tried to analyze relationships among teaching materials. However expected results were not derived, we changed the method to the TF-IDF method which focuses terms appearance. Finally we derived the general tendencies.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,500,000	450,000	1,950,000
2009 年度	1,100,000	330,000	1,430,000
2010 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：総合領域
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：オントロジー、RDF、OWL、Java プログラム、ソフトウェア解析

１．研究開始当初の背景

大学で作られている多くの教材資料は、著作権の問題ゆえにインターネットに公開されていない。そのため、学生にとって有益な

情報である可能性が高いにもかかわらず、インターネット用の検索エンジンでは検索できない。たとえ利用できたとしても、授業資料を構成する各ページが扱うトピックが持っている論理的な相互関係がわからず、その相互関係が数値化されていないために類似度などもわからない。せっかくの授業資料も再利用性が乏しいものとなっている。

２．研究の目的

研究担当者が所属する大学の学部で公開されているカリキュラム、シラバス、講義・演習資料、および一般から入手できる資料を対象として、カリキュラムと科目の関係、および教材資料に現れる概念用語間の関係と概念が記述されたページを関連付けるオントロジーを構築し、相互の関係や関係を利用した検索が行えるようなシステムを実現することを目指した。さらに、オントロジーをつかって教材間の定量的な関係を導出することを試みた。

３．研究の方法

（１）研究のサブゴール

授業用教材は、パワーポイント(以下PPT)、HTML、Excel、PDF の形式で書かれており、学部ホームページからたどってブラウズできるようになっている。これらの教材に現れる用語はあいまいであり論理関係のはっきりしないものが多い。そこで、あいまいさのない論理関係の明確なプログラムを扱う研究を並行して進めて基本的な技術とシステム開発を進めることとした。

このような方針からつぎのサブゴールを設定した。

①教材の収集と形式変換
②Java プログラムに関するオントロジー構築と推論・検索
③授業資料のオントロジー構築と推論・検索
④統合システムの開発
⑤教材間の関連の導出

（２）教材の収集と形式変換

学部ホームページには、カリキュラム、時間割、各授業のホームページだけでなく、各授業ホームページには各週の教材が公開されている。大学の方針により、授業が 2～3 年前から講義中心から演習を交えたものに変わってきたため、コンピュータで解析できる PPT や HTML 形式のものが減った。HTML 形式のものは統一的形式がないので除外し PPT 形式のもの約 50 科目を解析対象とした。PPT 形式のものを PPT2007 形式にして保存した。PPT2007 形式のものを解凍し XML 形式で保存されているデータから XSLT 変換して、科目名、開講週番号、ページ番号、ページタイトル、ページコンテンツ文字列からなる XML コンテンツを作成した。この XML コンテンツをネットワーク経由で日本語解析ソフト KoBaKo を使って形態素解析して XML 形式にした。さらにこの XML 形式となった解析結果から名詞を抽出して、RDF 形式に変換した。このとき用いた KoBaKo の辞書はデフォルトのものであり、普通名詞が多く現れ特殊な用語が登録されていないため、辞書の改良が必要なことがわかった。

（３）Java プログラムに関するオントロジー構築と推論・検索

用語の関係が論理的で、大量に入手できる Java プログラムを対象として、オントロジー構築と推論・検索の手法を開発する研究を並行して行った。デバッグの完了している Java バイナリコードを解析して、継承、関連を RDF 化しオントロジーを構築した。そののち、予想される結果が分かっている性質を推論してクラス間の関係を Sparql 文で検索した。予想通り良好な結果が得られた。つぎに規模を大きくして性能を解析した。データベースとして有名な約数万行の Hsqldb や H2DB の解析には 2 日間もの時間を要した。

設計思想を使えばよいのではと思い、設計図面である UML 図に設計思想が反映されているので、UML 図を対象とする研究へと展開した。その結果、検索文を図面で書きプログラムの UML 図を検索することが可能になった。この場合にも検索対象の UML 図を大きくすると計算量が発散した。

計算量爆発の原因を分析するとクラスとクラスの関連に関する推移律が原因であることがわかった。論理的には推移律は関連すべてに適用しなければいけないが、論理性があいまいな場合には推移律は２～３段の適用でよいと推測された。

（４）授業資料のオントロジー構築と推論・検索

最初のサブゴールの遂行において、教材解析の問題点は用語辞書にあることがわかった。そこでソフトを作って用語辞書作成に注力した。インターネットにある約 500 サイトの辞書サイトが登録されている Metapedia というサイトからメディア学部授業に関連する約 40 サイトの辞書ページから辞書用語を収集した。このデータを KoBaKo 辞書形式に変換して KoBaKo に登録した。これを用いて再度 PPT 教材を RDF 化した。これと並行して行っていた研究で問題となっていた複合語が単一用語として扱えるようになったが、分野が違えば用語が異なるということが問題であった。セマンティック Web では用語の同値を定義することで解決できる。しかし、これをコンピュータで自動的に行うことはできな

い。
教材の中には、その分野の概念を木構造で記述したものがあった。そこで、この木構造を RDFS で記述するとあいまいな概念用語の検索が可能になると予測されたので、本研究とは直接関係がないが、料理の分野における食材と調理法の RDFS を記述して推論・検索を行った。この場合には計算量の発散もなく、思い通りの結果（たとえば、肉野菜炒めの検索で牛肉キャベツ炒めが得られる）が得られた。その結果、教材にはその分野で必要となる概念用語の体系木、表現の異なる用語の関係などを記述することがよいことが分かった。

（５）統合システムの開発

各サブゴールに分けた研究では、処理をスクリプトで実行してきた。これでは実用にはならない。そこで、全体を統合するサーバ"meiseki"を開発し、研究を統合するとともに、ネットワークから検索ができるようにした。このシステムは、HPlab 開発の Joseki などと違い設定が極めて簡単で、しかも Web アプリケーションを書くためのタグライブラリを備えている。

学部のカリキュラム、シラバス、時間割などの情報からオントロジーを構築し、検索アプリをタグライブラリで実現し、学内からブラウザ経由で検索可能なシステムを実現した。

（６）教材間の関連の導出

教材に出現する用語を抽出しオントロジーを構築する際には、用語間の関係の記述が必要となる。KoBaKo の用語辞書を強化しても、あまり改善がみられなかった。そこで、用語の出現頻度に着目する TF-IDF 法を用いて科目間の関連の強さを定量化することとした。この方法で約 50 科目同士の関連を定量化した。その結果、当学部の 3 分野（技術、表現、環境）のうち技術分野は科目間に関連の深い数値がえられたが、環境分野では同一教員による授業以外には関連があまりなかった。表現分野では PPT 資料が解析対象となるような形式のものが少なかった。

（７）今後の展開

セマンティック Web を推進する W3C では、セマンティック Web 情報を HTML に埋め込ませる規格 RDFa を作り、HTML ページを書かせようとしている。この動向をみならい、セマンティック Web 情報を PPT に埋め込んで推論・検索を容易にする研究に着手した。

つまり、PPT の最後のページなどに、重要用語を列挙し、重要用語やページの関連を木構造などの形で記述してもらう。その構文チェックをすることにより正確な記述を担保できる。

４．研究成果

（１）既存教材に関して

①HTML/PPT/XML 形式で公開されている情報から XML 形式に変換する XSLT パッケージを作成した。このパッケージを用いることにより、PPT、HTML、Excel など様々な形式のデータを簡単に XML 形式に変換し、RDF 化して保存する技術が完成した。

②PPT 形式の教材でセマンティック Web 技術を使えるようにするためには、既存の PPT では用語と用語間の関係がはっきりしないことが問題であることが分かった。それを解決するには、PPT にセマンティック情報、つまり重要用語の記述と用語間の関係を木構造などで記述することが必要であることが分かった。この研究は、現在、修士の研究テーマとして実施している。

（２）Java 技術に関して

①Java 言語で記述されたプログラムのコンパイル結果からプログラム間の関係をセマンティック Web 技術で推論する技術ができた。これを学会で発表後、学会の１つのセッションが設立された。

②RDF グラフを作成する際に利用するツールに対応する UML ツールを使って図面で書かれた情報から図面で検索する技術ができた。これも学会で強い関心を持たれている。

（３）統合技術に関して

①上記の一連の作業を自動化し、ネットワークに Web サーバ meiseki を開発した。このシステムは、従来広く使われてきた HP 研究所開発の Joseki に代わって設定ファイルが XML 形式になっており簡単に利用できる。Sparql に関する書籍の例題約 200 のいずれもが正しく動作することが確認でき、完成度が非常に高いことが分かった。また、従来システムではアプリケーションを Java 言語で書かざるをえなかったが、このシステムではラグライブラリを開発したため、JSP と JSTL を用いてだれでもが簡単に作成できるようになった。

（４）教材の定量化に関して

①教材のうち、技術関係は用語が定着しており人による揺らぎは少ない。それに対して非技術分野は人によって用いる用語が違ううえ複合語が多く、用語の揺らぎが少ないことを想定したオントロジー構築は難しい。そのため、オントロジーを前提とした定量化は難しい。用語の出現頻度などを前提とした解析しかできなかった。

（５）まとめ

①この研究を支える基礎技術、統合システム技術に関しては完成度の高いものができた。教材そのものに関しては、内部で用いられている用語の表現が揺らいでおり思うような結果は得られなかった。しかし、別のシステムを作って実験した結果、これを解決するに

は、教材を作成する人に用語と用語間の関係を書かせることが必要であることが分かった。

このことは、教材作成にあたっては、常に授業の全体を考慮して各週の教材と各教材の各ページを論理的なものにさせることにより、授業全体の改善に資するものと思う。PPT を作成する人にこのようなことを強要することは、HTML に関して W3C が進めている HTML に記述形式を強要する RDFa と一致している。

なお、この研究をベースとして、学部 3 年生を対象に学部の教材ホームページをセマンティック Web 化する 1 学期間の演習を実施中である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0 件）

〔学会発表〕（計　13 件）
①青木祐香里、塚本享治：“meiseki”セマンティック Web サーバの開発、第 10 回情報科学技術フォーラム(2011.9.7) 函館大学（北海道）
②長谷川明史、塚本享治：クラス図とシーケンス図からのセマンティック Web 技術を用いたソフトウェアの検索とその評価、情報処理学会第 171 回ソフトウェア工学研究会(2011.3.7) 化学会館(東京都)
③青木祐香里、塚本享治：スクリーン・スクレーピングを用いたビジネス統合方法、第 73 回情報処理学会全国大会(2011.3.3) 東京工業大学（東京都）
④長谷川明史、塚本享治：クラス図とシーケンス図からのセマンティック Web 技術を用いたソフトウェアパターン検索、第 73 回情報処理学会全国大会(2011.3.3) 東京工業大学（東京都）
⑤長谷川明史、塚本享治：シーケンス図をクエリとして用いるシーケンス図の振舞検索の提案、電子情報通信学会知能ソフトウェア工学研究会(2011.1.25) 機械振興会館（東京都）
⑥長谷川明史、塚本享治：クラス図によるクラス図構造検索の評価、第 9 回情報科学フォーラム(2010.9.7) 九州大学（福岡県）
⑦長谷川明史、塚本享治：クラス図をクエリとして用いるクラス図構造検索手法の提案、情報処理学会第 169 回ソフトウェア工学研究会(2010.7) 北九州テレワークセンタ（福岡県）
⑧長谷川明史、塚本享治：UML で記述されたソフトウエアの RDF グラフへの変換、情報処理学会第 72 回全国大会 6P-5 (2010.3.11)東京大学(東京都)
⑨長谷川明史、西村紅美、塚本享治：セマンティック Web 技術を用いた PC パーツの検索、情報処理学会ソフトウェア工学研究会 2010-DD-74 (2010.1.29) 甲南大学(兵庫県)
⑩北村怜子、塚本享治：プロパティの階層化を用いた教材資料からのオントロジー構築、情報処理学会第 72 回全国大会 2Z-5 (2010.3.9)東京大学(東京都)
⑪長谷川明史、塚本享治：OWL2.0 の推論を用いたオープンソース Java ソフトウェアの構造検索と評価、情報処理学会ソフトウェア工学研究会(2009.3.18)情報処理学会会議室(東京都)
⑫長谷川明史、塚本享治:セマンティック Web 技術を用いた Java ソフトウェアの構造検索、情報処理学会第 71 回全国大会(2009.3.13) 立命館大学(滋賀県)
⑬北村怜子、塚本享治：教材資料を対象とした用語収集とその分析手法、情報処理学会第 71 回全国大会(2009.3.12)立命館大学(滋賀県)

６．研究組織
(1)研究代表者
塚本 享治（TSUKAMOTO MICHIHARU）
東京工科大学・メディア学部・教授
研究者番号：90386764

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：