様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月　６日現在

機関番号：６２６１５
研究種目：若手研究（B）
研究期間：２０１０～２０１１
課題番号：２２７００１６０
研究課題名（和文）　データ統合のための意味的な関係知識の発見技術に関する研究

研究課題名（英文）　Relation Discovery for Semantic Integration

研究代表者
市瀬　龍太郎（ICHISE RYUTARO）
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：００３３２１５６

研究成果の概要（和文）：本研究では，さまざまなデータを動的に統合し，ユーザに必要な情報を届けることができるデータ基盤を構築する技術を開発する．そのために，インターネット上に大量に存在する様々な結合データ同士の意味的な関係を自動的に発見することができる新たなデータ統合技術の開発を行った．

研究成果の概要（英文）：The purpose of this study is to develop a technology to construct data infrastructure that can dynamically integrate various data with semantics, and deliver it to users. In order to do it, we developed a novel relation discovery method for semantic integration.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,800,000	540,000	2,340,000
2011 年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知能情報処理

１．研究開始当初の背景

Ｗｅｂには，様々な情報が掲載され，それらの情報が人々の意思決定に大きな役割を果たしている．しかし，現在のＷｅｂは，意思決定に必要な情報を人手で収集，統合しなければならない．そのような作業を自動化するため，セマンティックＷｅｂにおいて，様々なデータを結合データ（Linked Data）と呼ばれる形式で公開する試みが行われている．結合データは，辞書的な語のデータや地理のデータ，アーティストの人物データなど様々なデータを特定のデータ形式で表したものであり，それらをお互いに関係づけることで，意味的な情報処理を機械に行わせる基盤としている．これにより，分散したデータを透過的に扱うことが可能となる．この試みは 2007 年頃から始ま

り，2009 年春の時点で，42 億個のデータが公開されるなど，急速に広まっている．しかし，これらのデータは，各々，独立に作られているため，意味のある情報検索，情報統合を行わせるには，お互いの結合データ同士を意味的に関連付けなければならない問題点がある．本研究では，この問題を結合データ連結問題と呼ぶ．たとえば，駅名データベースに入っている「東京」は，都道府県データベースに入っている「東京」と同じ文字列であるが，位置情報としては包含関係になっており，このような意味的な関係づけが無い限り，知的な処理を行わせることは難しい．そのため，大量の結合データに対して，お互いの意味的な関係づけを自動的に行う手法の開発が喫緊の課題となっている．

国際的にも，データ間の意味的な関係付けをいかに行うかという研究が多くなされており，特にセマンティックＷｅｂの研究ではその動きが顕著である．セマンティックＷｅｂにおいて，意味を取り扱うためには，オントロジーを用いるものとされた．そのため，オントロジー・アライメント問題の解決は，大きな焦点となった．この問題に対し，2004 年より共通データセットによるコンペティション形式の国際性能評価研究会が毎年開かれ，多数の参加者を得ている．しかし，2007 年に結合データの概念が提起されると，セマンティックＷｅｂの研究では，結合データが研究の中心の一つとなり，多くの結合データが公開されるに従って，結合データ連結問題が大きな課題となってきた．

２．研究の目的

本研究の目的は，大量の結合データに対して，お互いの意味的な関係づけを自動的に行う手法の開発である．そこで，本研究では，次を実施することによって，研究の目的の達成を試みる．

・言語の類似性などを利用したオントロジー・アライメント技術と，大規模なグラフを取り扱うリンク・マイニング技術の組み合わせによる結合データの高精度な自動関係付け手法の開発

目的の達成のためには，研究代表者が，これまでに開発してきた，機械学習に基づくオントロジー・アライメント手法に，大量のデータを取り扱うことができるリンク・マイニングの手法を統合していくことが，これまでの研究より有効であると考えられる．そのため，これらを有機的に融合した新たな手法を開発することで，結合データ連結問題の解決を図る．

３．研究の方法

本研究では，言語の類似性などを利用したオントロジー・アライメント技術と，大規模なグラフを取り扱うリンク・マイニング技術を組み合わせることにより，結合データの高精度な自動関係付けする機構を新たに開発し，結合データ同士を意味的に関連付けする問題の解決を図った．

２年間の研究は，下記のような形式で遂行した．

（１）　２０１０年度

２０１０年度は，研究の初年度に当たるため，主に研究に必要な環境の整備に焦点を当て，下記の２つに分けて研究開発を実施した．

① 結合データの収集と研究用大規模データセットの開発

本研究を開始するに当たり，あらかじめ，分散して存在する結合データを収集し，手元の計算機において，研究ができるような環境の整備を行う．そのための結合データの収集，および，結合データを容易に利用できるようなデータベース環境，データセットの開発を行った．

② オントロジー・アライメント手法，リンク・マイニング手法の適用による意味関係の解析

これまでに，開発してきた機械学習技術に基づくオントロジー・アライメント手法，リンク・マイニング手法を適用することによって，データの意味に関する特性を明らかにした．これにより，オントロジー・アライメント手法，リンク・マイニング手法を組み合わせた効果的な手法を開発するための基礎データを得た．

（２）　２０１１年度

２０１１年度は，２０１０年度

に整備した研究環境を用いた解析結果に基づき，以下の２つに分けて研究開発を行った．

① オントロジー・アライメントとリンク・マイニングを融合した精度の高い結合データの意味的な関係づけ手法の開発

本研究では，データ間の意味的な関係づけをおこなうために，クラスターベースの類似度統合システムを作成した．実験した結果，従来手法よりも高い精度で，意味的な関係を抽出できることが示され，情報同士を精度高く意味的に統合できることが可能となった．

② 大量の結合データへの対処方法の開発

大量の結合データを取り扱うために，本研究では結合データから必要な部分のみを抽出して，意味的な統合を行う手法を開発した．本手法により，取り扱うデータ量を大幅に減らしながら，結合データの統合を行うことが可能となった．

４．研究成果

（１） ２０１０年度

前章で述べた研究で得られた成果に基づき，論文３本の出版を行った．以下，それぞれの論文の概要を記述する．なお，冒頭の番号は,「５．主な発表論文等」に記載されている論文番号と対応している．

④ 単語同士の意味的な類似度を計測するための新たな指標を提案し，実験的に評価を行った．

⑤ 既存のデータベースから，結合データを自動的に生成する機構を開発した．

⑥ 様々な類似度尺度に対して，機械学習手法などを利用して分析を行った．

（２） ２０１１年度

前章で述べた研究で得られた成果に基づき，論文３本の出版を行った．以下，それぞれの論文の概要を記述する．なお，冒頭の番号は,「５．主な発表論文等」に記載されている論文番号と対応している．

① さまざまな類似尺度を統合するためのクラスターベースの手法を提案し，実験的に評価を行った．

② 結合データを意味的に統合するために，オントロジーを自動構築して利用することを提案した．

③ 大量の結合データの中から，必要な部分のみを抽出することによって，オントロジーの自動構築を効率化する手法を提案し，実験的に評価を行った．

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６件）

① Quang-Vinh Tran, Ryutaro Ichise, Bao-Quoc Ho: Cluster-based Similarity Aggregation for Ontology Matching, Proceedings of the 6th International Workshop on Ontology Matching, pp.142-147, 2011,(査読有)

② Lihua Zhao, Ryutaro Ichise: One Simple Ontology for Linked Data Sets, Proceedings of the ISWC 2011 Poster and Demonstrations Track, 2011,(査読有)

③ Lihua Zhao, Ryutaro Ichise: Mid-Ontology Learning from Linked Data, Proceedings of Joint International Semantic Technology Conference, pp.112-127, LNCS 7185, 2011,(査読有)

④ Raul Ernesto Menendez-Mora, Ryutaro Ichise: Effect of Semantic Differences in WordNet-Based Similarity Measures, Proceedings of the 23rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Vol. 2, pp. 545-554, LNAI 6097, 2010,(査読有)

⑤ Simeon Polfliet, Ryutaro Ichise: Automated Mapping Generation for Converting Databases into Linked Data, Proceedings of the ISWC 2010 Poster and Demonstrations Track, pp. 173-176, 2010,(査読有)

⑥ Ryutaro Ichise: An Analysis of Multiple Similarity Measures for Ontology Mapping Problem, International Journal of Semantic Computing, Vol. 4, No. 1, pp. 103-122,

2010,(査読有)

６．研究組織

(1)研究代表者

市瀬　龍太郎（ICHISE RYUTARO）
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：００３３２１５６

(2)研究分担者

該当なし

(3)連携研究者

該当なし