Vol.2014-IOT-27 No.20
Vol.2014-SPT-11 No.20
2014/10/10

オントロジを用いたLinked Dataの構造理解支援システムのモデルの提案

高屋敷 健[1]　児玉 英一郎[1]　王 家宏[1]　高田 豊雄[1]

概要：近年，Linked Data に関する研究が活発になっている．実際，日本においても，Linked Data のリンク修復に関する研究や，Linked Data を利用したアプリケーション開発に関する研究などが数多く行われている．Linked Data を利用する際には，SPARQL クエリの作成が必要となるが，この SPARQL クエリの記述のためには，Linked Data におけるリンク構造の理解が必要となる．しかし，Linked Data におけるリンク構造の理解は，一般的には困難なものとなっている．そこで，本研究では，オントロジを用いた Linked Data の構造理解支援システムのモデルの提案を行う．本モデルでは，オントロジを用いることにより，SPARQL クエリ作成の際に必要な，実際に使用可能であるリソース，プロパティ，プロパティ値を検索することが可能となっている．また検索結果の可視化により SPARQL クエリ作成をさらに支援している．本研究では，本提案モデルの有用性に関する評価結果についても報告する．

A Ontology-based Support System for Understanding Structure of Linked Data

Takeru Takayashiki[1]　Eiichiro Kodama[1]　Jiahong Wang[1]　Toyoo Takata[1]

Abstract: In recent years studies on Linked Data have become active. In Japan, studies on link restoration of Linked Data and development of applications using Linked Data have been reported. When we use the Linked Data, generally we have to write queries using the SPARQL query language. And for writing SPARQL queries, it is necessary to understand the link structure of Linked Data. The problem is that, however, understanding of the link structure of Linked Data is difficult in general. This paper proposes an ontology-based support system model for understanding the link structure of Linked Data. To support query writing, using the ontology, the proposed system model makes it possible to search for available resources, properties and property values. The query writing is further supported with the visualization of searching results. This paper also reports the evaluation results about the usefulness of the proposed model.

1. はじめに

近年，セマンティック Web の分野において Linked Data[1] の研究が盛んに行われている．Linked Data とは，RDF(Resource Description Framework) 形式で公開されたリンクするデータを指しており，様々な領域において作成されている．この Linked Data に関しては，公開された Linked Data を収集，蓄積し，データの利用の促進を図る LOD(Linked Open Data) プロジェクト [2] が知られている．

Linked Data に関する研究は，特に欧米にて盛んに研究されており，その成果は多くの企業などで利用されている．一例として，BBC による Linked Data を利用した歌手情報の提供が知られている．日本においても Linked Data に関する研究は活発になっており，Linked Data のリンク修復に関する研究や，Linked Data を利用したアプリケーション開発に関する研究などが行われている．しかし，Linked Data に関する研究が増加する一方で，企業などによる Linked Data を利用したアプリケーションの開発事例は少ない状況にある．Linked Data を公開している

[1] 岩手県立大学大学院ソフトウェア情報学研究科
Greduate School of Software and Information Science, Iwate Prefectural University

Vol.2014-IOT-27 No.20
Vol.2014-SPT-11 No.20
2014/10/10

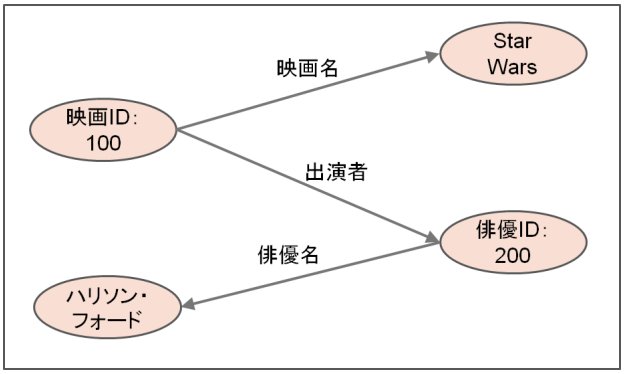

図 1　映画のデータセット内のリンク構造の例

データセットごとに，Linked Data 内に記述された情報の表現方法が異なっていることや，また，データセット特有のプロパティが利用されていることが原因で，SPARQL クエリの作成が困難であるためだと考えられる．

本研究では，SPARQL クエリの作成困難な部分の解消を図り，日本における Linked Data の利用の促進を目的として，オントロジを用いた Linked Data の構造理解支援システムのモデルの提案を行う．

2. Linked Data

2.1 Linked Data

Linked Data は 2006 年に Tim Berners-Lee によって提唱された，外部のデータとリンクしている RDF 形式のデータのことである．RDF[3] とは Semantic Web の分野におけるメタデータを記述するためのフレームワークのことであり，一般の Linked Data は，RDF に従った複数のトリプルから構成されている．トリプルは，Linked Data の構成単位であり，リソース (主語)，プロパティ (述語)，プロパティの値 (目的語) の 3 つの要素からなっており，リソースの関係情報を表したものとなっている．例として，「日本の首都は東京である」をトリプルで表現すると，リソースが日本，プロパティが首都，プロパティの値が東京となる．

公開されている Linked Data として Wikipedia を Linked Data 化した DBpedia[4] や映画のデータセットである Linked Movie DataBase（LinkedMDB）[5] などが知られている．図 1 に LinkedMDB 内に記述されている Star Wars の出演俳優名を指すトリプルを有効グラフにて示す．図 1 は，映画のデータセット内の映画名，Star Wars から出演俳優名までのリンク構造を表現している．

また，DBpedia のような Linked Data は，一般的に RDF ストレージと呼ばれるデータベースへ格納することにより利用する．RDF ストレージの代表的なものとしては，Virtuoso[6] や Sesame[7] などが知られている．

```
PREFIX linkedmdb:
<http://data.linkedmdb.org/resource/>
SELECT DISTINCT ?actorName WHERE {
?film linkedmdb:film_name “Star Wars".
?film linkedmdb:movie/actor ?actor.
?actor linkedmdb:actor_name ?actorName.
}
```

図 2　SPARQL クエリ記述例

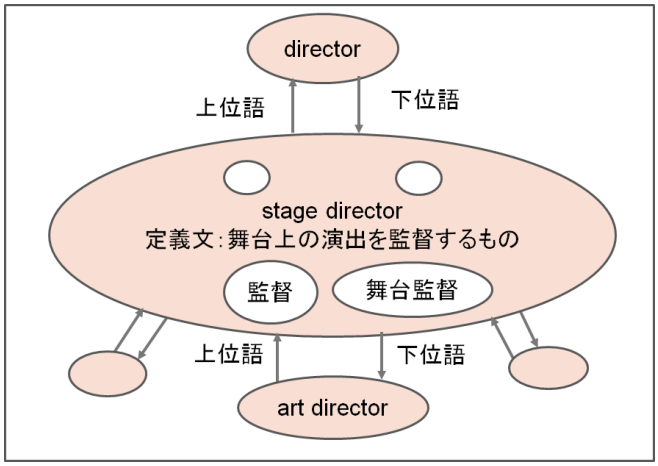

図 3　日本語 WordNet

2.2 SPARQL

SPARQL(SPARQL Protocol and RDF Query Language)[8] とは，RDF ストアへ格納された RDF 形式のデータに対し，検索，操作を行うための言語である．Linked Data を公開している Web サイトでは，SPARQL を利用したクエリに応答するサービスを，SPARQL Endpoint として公開している．図 1 で示した映画のデータセット内から Star Wars の出演俳優名を取得する SPARQL クエリの例を図 2 に示す．

図 2 における SPARQL クエリによって，図 1 におけるハリソン・フォードのような俳優名が取得できる．この例のように，SPARQL クエリの作成においては，Linked Data のリソースやプロパティの値がプロパティによってどのようにリンクされているかというリンク構造を理解して記述する必要がある．即ち，SPARQL クエリ作成の際には，事前準備として，Linked Data のリンク構造の理解や，使用されているプロパティを知ることが必要となる．

3. 日本語 WordNet

日本語 WordNet[9][10] は，NICT が開発した日本語の概念辞書 (オントロジ) であり，プリンストン大学で開発された英語オントロジである WordNet の日本語版となっている．この WordNet では，単語を，synset と呼ばれる概念を表すグループに分類している．例えば監督という単語

は stage director という synset に含まれている．synset には，語や synset 間の関係が記述されており，synset 間の上位，下位関係や，語の間の同義語関係が取得可能である．図 3 に synset の語や synset 間の関係の例を示す．

図 3 は，stage director の synset の例となっている．stage director には，上位の synset として director があり，下位の synset として art director などがある．また，stage director 内には，synset の概念を説明した定義文や，同義語である監督や舞台監督などが記述されている．

4. SPARQL クエリ作成における問題点

前述のように，Linked Data を利用したアプリケーションは，現在のところ少ない状況にある．これは，Linked Data の利用においては，SPARQL クエリを作成しなければならないためであると考えられる．新たに利用する Linked Data に対し，SPARQL クエリを作成する場合には，リンク構造を理解しなければ，作成することはできない．しかし，Linked Data の構造の理解のための方法は，現状，限られている．Linked Data の構造の理解をするための方法として，次の 3 つが考えられる．まず第 1 の方法として，データセットが Linked Data を Web ページとして公開している場合に，Web ページ内のリンクを辿り理解する方法である．しかし，Linked Data を Web ページとして公開しているデータセットは多くはない．

第 2 に，RDF ファイルを直接調査し，構造を理解する方法がある．しかし，トリプル数が数百万単位になる Linked Data も数多くあり，ファイルを直接調査する方法は現実的ではない．

最後に SPARQL クエリを利用してリンク構造を理解をする方法が考えられる．検索要求に関連する語をリソースやプロパティの値として持つようなトリプルを SPARQL クエリによって出力し，少しずつ理解を行う方法である．しかし，少しずつクエリの改良を行いながらリンク構造の理解を行うため，時間や手間を要する．

このようなことから，Linked Data の構造理解の支援を行うことができれば，SPARQL クエリ作成の支援となり，Linked Data を利用したアプリケーション作成の一助になると考える．

5. 関連研究

5.1 Linked Data Query Wizard

Linked Data の利用の支援を行う関連研究として，Linked Data Query Wizard[11] が知られている．Linked Data Query Wizard では，検索語を用い，Label というプロパティの値のなかに検索語を部分文字列として含むようなプロパティの値と，そこで利用されているプロパティを取得することができ，検索結果は表形式にてユーザへ提示される．また，表において別なプロパティを選択することにより，そのプロパティの値が表中へ出力される．さらに，検索結果のプロパティの値と追加したプロパティの値の関係の可視化も行える．

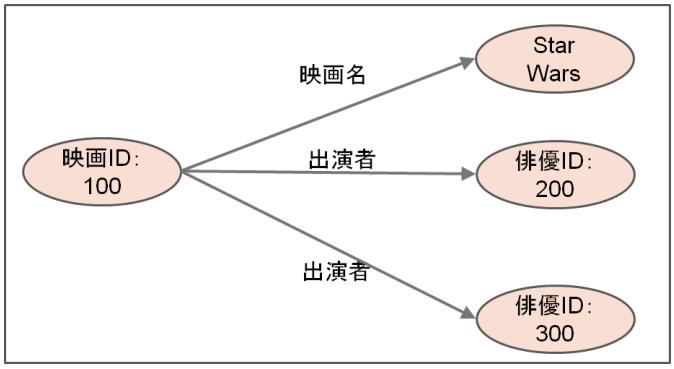

図 4 関連研究における可視化の例

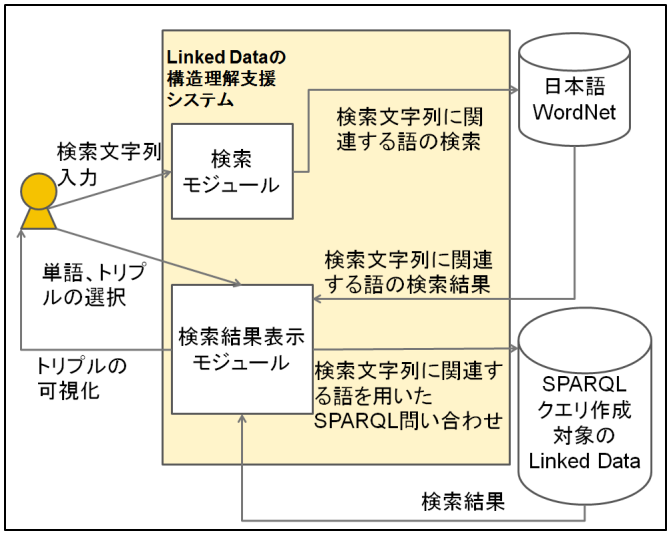

図 5 オントロジを用いた Linked Data の構造理解支援システムのモデル

5.2 関連研究の問題点

関連研究の問題点として，プロパティの意味が分からない場合には，そのプロパティを表に追加するべきか，すべきでないか判断できないことが挙げられる．また，関連研究における可視化は，図 4 のような検索語がリソースとなっている場合のトリプルのみであり，図 1 のようなプロパティの値がほかのトリプルのリソースとなっているようなリンク構造のものは提示できない．このため，複数のトリプルを利用する SPARQL クエリを作成する場合，映画 ID から俳優名までのリンク構造のような複数のトリプルをまたいだ関係の記述に際し，支援が不十分となっている．

6. 提案手法

本研究で提案するオントロジを用いた Linked Data の構造理解支援システムのモデルを図 5 に示す．以下，図 5 内の各構成要素について，順に説明を行う．

Vol.2014-IOT-27 No.20
Vol.2014-SPT-11 No.20
2014/10/10

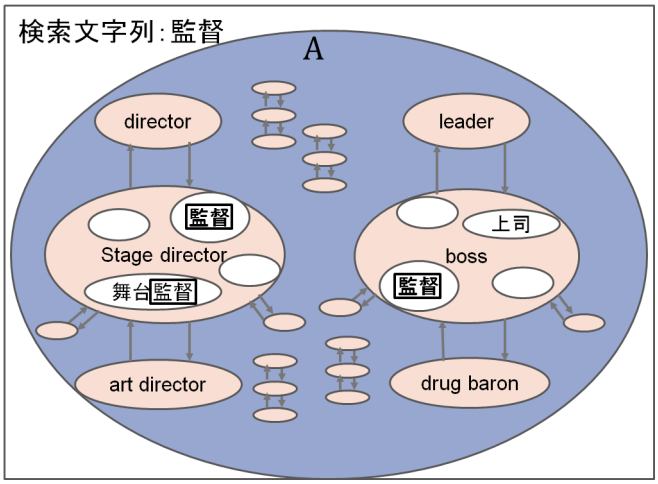

図 6　オントロジを利用した検索で作成される和集合 A の例

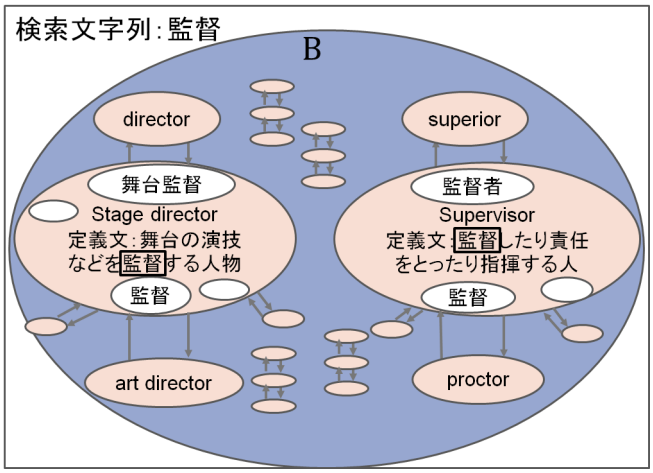

図 7　オントロジを利用した検索で作成される和集合 B の例

6.1　Linked Data の構造理解支援システム

Linked Data の構造理解支援システムでは，オントロジを利用し検索文字列に関連する語の検索を行い，その関連する語がプロパティやプロパティの値として利用されているか，SPARQL クエリ作成対象の Linked Data に対し問い合わせを行う．また，問い合わせ結果として取得したプロパティやプロパティの値を利用し，トリプルの可視化を行い，リンク構造の理解の支援を行う．以下，提案モデルの流れを示す．

(1) 入力された検索文字列を用いて日本語 WordNet を利用し，検索文字列が含まれている単語が属している全ての synset と，それぞれの上位，下位の synset の取得を行い，その和集合 A を作成する．

(2) (1) と同様に日本語 WordNet を利用し，検索文字列が含まれている定義文が属している全ての synset と，それぞれの上位，下位の synset の取得を行い，その和集合 B を作成する．

(3) $A \cap B$ に属する単語があった場合，その単語をユーザに対し提示する．

(4) ユーザによって選択された単語を用い，SPARQL クエリ作成対象のデータセットに対し，SPARQL クエリを用いた問い合わせを行い，単語が含まれているプロパティとプロパティの値を取得する．

(5) ユーザにより選択されたトリプルとそのトリプルに隣接しているトリプルの表示を行う．

(6) 終端のトリプルが選択された場合には，その先のトリプルを提示する．

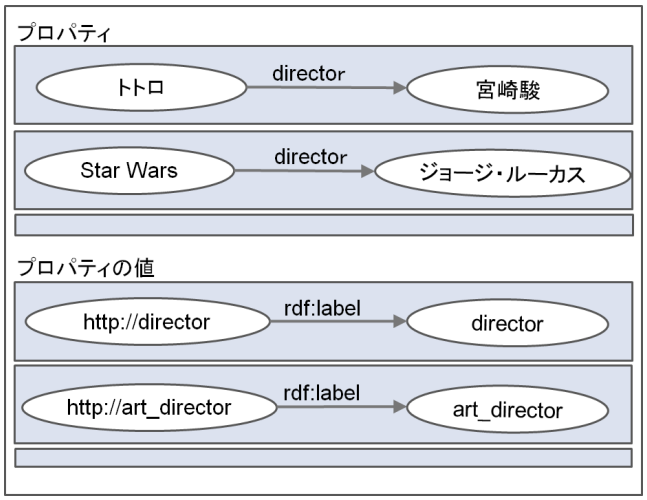

図 8　オントロジを用いた検索結果の表示例

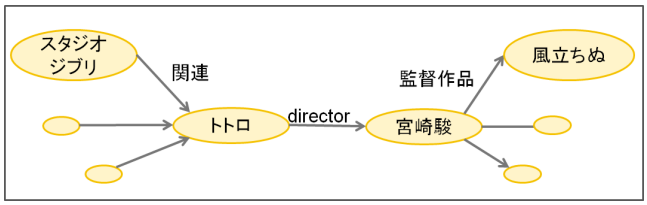

図 9　トリプルの可視化の例

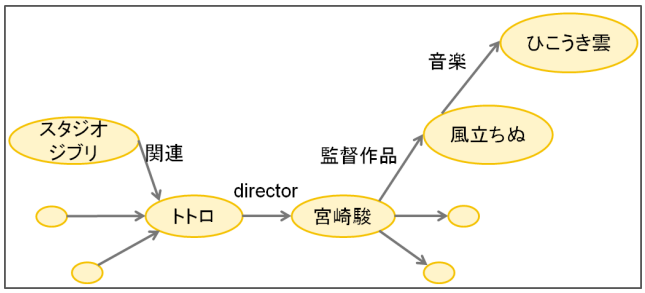

図 10　可視化されたトリプルの追加の例

6.2　本提案システムの利用例

以下，上述の流れを具体的な例を用いて説明する．本提案モデルが使用される状況として，ユーザが監督名からその監督の作品，及び，その作品内で使用されている音楽を検索する SPARQL クエリの作成を行うことを想定する．

ユーザにより本提案モデルに対し，検索文字列として「監督」が入力された場合，手順 (1) と (2) に従い図 6 と図

7 のような和集合 A，B が作成される．

次に $A \cap B$ に属する単語，「監督」や「美術監督」，「director」などの検索文字列に関する単語がユーザに提示される．提示された単語群の中から，ユーザは SPARQL クエリ作成対象のデータセットに対し問い合わせを行うための単語の選択を行う．例えば，ユーザが「director」を選択した場合，SPARQL クエリによる問い合わせ結果は，図 8 のように提示される．

図 8 から「となりのトトロの diretor は宮崎駿である」というトリプルを選択した場合，図 9 のように可視化される．

また，可視化されたトリプル内の終端のトリプル「宮崎駿の監督作品は風立ちぬである」を選択した場合，図 10 のように選択したトリプルの先のトリプルが表示される．

7. 評価

本提案システムの評価では，関連研究との比較のため関連研究の評価にて利用されている NASA Task Load Index[12] を利用する．

7.1 NASA Task Load Index(NASA TLX)

NASA TLX はあるタスクに対する負荷仕事量 (workload) を示すための手法である．タスクの仕事量を表現するために，基本となる 6 つの評価尺度が用意されており，この 6 つの評価尺度ごとに 50 を標準値とした 0〜100 の数値によって，被験者による主観的評価をつけ，タスクの仕事量を算出するものとなっている．算出された仕事量は低いほど良い．以下に 6 つの評価尺度の詳細を示す．

・Mental Demand(MD)

タスクに対し，計算や決定などの，知的な要求が多かったか，少なかったかの評価を行う．

・Physical Demand(PD)

タスクに対し，マウス操作やキーボード操作などの身体的活動が，過酷であったか，過酷でなかったかの評価を行う．

・Temporal Demand(TD)

タスクを行った際に感じた，時間的切迫感が高かったか，低かったかの評価を行う．

・Effort(EF)

タスクを成し遂げるためにどれだけ一生懸命おこなったかについて，多くの努力を要したか，少ない努力で行えたかの評価する．

・(Own)Performance(OP)

タスクの目標をどの程度達成できたかについて，達成度が高いほど，低く評価を行う．

・Frustration(FR)

タスクを行った際に感じたストレスについて，高いか，低いかの評価を行う．

本研究では，RTXL(Raw TXL) と呼ばれる上記の 6 項目の平均を仕事量とする．

7.2 評価方法

評価は実験を行い，その結果から評価を行う．関連研究では，Linked Data に関する知識を持つ 14 人の被験者に対し評価実験を行っている．また，14 人の被験者のうち，7 人の被験者は Linked Data に関する知識を多少持ち合わせている程度の者であった．そこで，本研究の評価では，Linked Data について事前に学習を行ってもらい，Linked Data に関してリンク構造を理解できる程度の知識を持った被験者 6 名に対し評価実験を行い，関連研究の知識量をある程度持っている 7 人の被験者との比較を行う．実験 1 では，関連研究と同じプロパティの値を対象とした検索を行う．また実験 2 では，プロパティを対象とした検索を行い，実験 1 と同様なリンク構造の発見を行う．具体的な実験内容を以下に示す．

・実験 1

(1) 映画「パルプ・フィクション」の映画監督はだれか．
(2) 映画「パルプ・フィクション」の映画監督の他の作品は何か．
(3) 映画「パルプ・フィクション」にはブルース・ウィリスは出演しているか．出演している場合，ブルース・ウィリスが出演している他の映画を探す．

以上の 3 項目について DBpedia の日本語版である DBpedia Japanese[13] に対し，制限時間を 5 分として評価を行った．

・実験 2

(1) 被験者が決定した映画監督の作品．
(2) (1) において発見した作品の出演俳優．
(3) (2) において発見した出演俳優が出演している他の映画．

以上の 3 項目について映画のデータセットに対し，制限時間を 5 分として評価を行った．

7.3 評価結果

全被験者の RTLX の平均値を表 1 に示す．また，実験 1 の結果を図 11，実験 2 の結果を図 12，関連研究において実験 1 を行った結果を図 13 に示す．

8. 考察

表 1 の RTLX の平均値は実験 1 と関連研究を比較した場合，本提案モデルのほうが RTLX の平均値が低く良い結果となった．しかし，実験 2 と関連研究を比較した場合，関連研究の方が RTLX の平均値が低く，良い結果を出している．これらの点について，以下，考察を行う．

実験 1 の結果である図 11 と関連研究の結果である図 13 を比較した場合，OP と FR 以外は大きく変わらないが，OP は比較対象よりも大きく下回る結果となっている．このことから，関連研究よりも Linked Data のリンク構造を

表 1　RTLX の平均値

実験 1	実験 2	関連研究
37.3	48.5	43.1

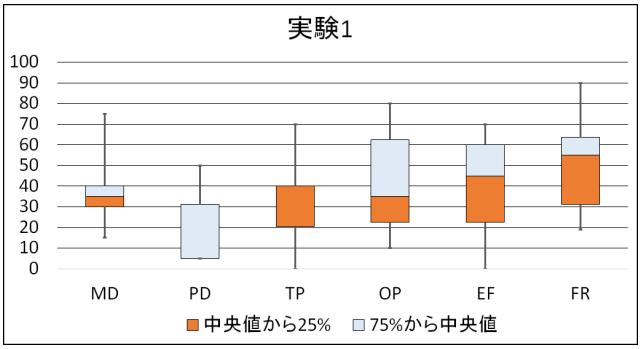

図 11　実験 1 の結果

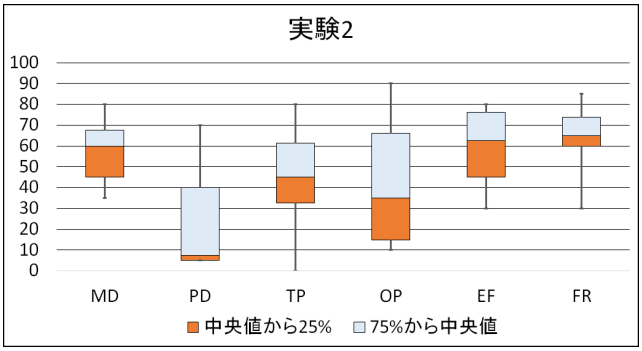

図 12　実験 2 の結果

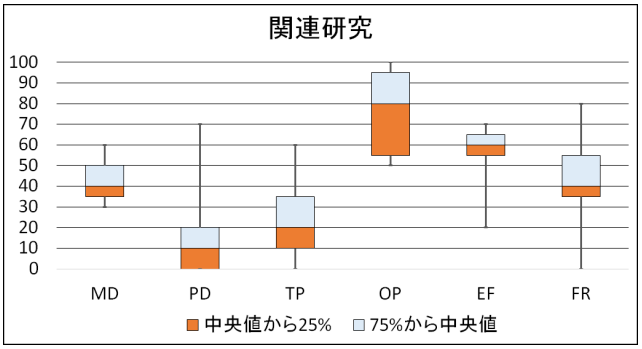

図 13　関連研究の結果 [11]

理解しやすいと考えられる．しかし，提案システムの可視化後のトリプルの追加作業により，可視化されるトリプル数が大幅に増えるため，目的のトリプルの捜索が困難になり，FR が高くなったと考えられる．

次に実験 2 の結果である図 12 と，関連研究の結果である図 13 を比較した場合，OP 以外の数値は関連研究よりも若干高い数値となった．これは，LinkedMDB が DBpedia よりもリンク構造が複雑なことにより，操作数が増えたためだと考えられる．しかし，作業成績である OP は本提案モデルのほうが低い数値となったことから，リンク構造を理解しやすいと考えられる．

9. おわりに

本研究では，Linked Data の構造理解支援システムのモデルの提案と，NASA TLX を用いた関連研究との比較評価を行った．プロパティの値を対象とした検索の比較評価では，関連研究よりも NASA TLX の数値で 5.8 ポイント低い結果を出すことができたが，プロパティを対象とした検索の比較評価では，関連研究よりも 5.6 ポイント高い結果となった．しかし，OP の値は関連研究よりも低いため，リンク構造の理解は行いやすいと考えられる．今後の課題として，可視化されたトリプルの中から目的のトリプルの発見を行いやすくするために，可視化されたトリプルに対するオントロジーを利用した検索機能の実装が考えられる．可視化されたトリプルに対するオントロジーを利用した検索を行うことで，目的のトリプルの発見に要する時間が短縮され，TP や EF，FR の値を下げることが可能であると考えられる．

参考文献

[1] Tim Berners-Lee: Design Issues: Linked Data, http://www.w3.org/DesignIssues/LinkedData.html, (2006)
[2] Linked Open Data, http://lod-cloud.net/
[3] RDF/XML Syntax Specification (Revised), http://www.w3.org/TR/REC-rdf-syntax/
[4] DBpedia, http://dbpedia.org/About
[5] Linked Movie Data Base, http://www.linkedmdb.org/
[6] Virtuoso, http://virtuoso.openlinksw.com/
[7] Sesame, http://www.openrdf.org/
[8] Andy Seaborne: SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
[9] Francis Bond, Timothy Baldwin, Richard Fothergill, Kiyotaka Uchimoto: Japanese SemCor: A Sense-tagged Corpus of Japanese, The 6th International Conference of the Global WordNet Association (GWC-2012), pp.1–8 (2012).
[10] Kow Kuroda, Francis Bond, Kentaro Torisawa: Why Wikipedia needs to make Friends with WordNet, The 5th International Conference of the Global WordNet Association (GWC-2010), pp.9–16, (2010).
[11] Patrick Hoefler, Michael Granitzer, Eduardo Veas, Christin Seifert: Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints, LDOW2014(WWW2014), pp.1–10 (2014).
[12] Hart, S. G., Staveland, L. E.: Development of NASA-TLX (Task Load Index) Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (eds.), Human Mental Workload, pp.139–183 (1988).
[13] DBpeida Japanese, http://ja.dbpedia.org/