人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第5回)
SIG-AM-05-02

RDF データベースを対象とした データ分析支援ツールの提案

Proposal of Data Analysis Support Tool for RDF Database

田代 航一* 高間 康史

Koichi Tashiro, Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

Abstract: 本稿では，RDF の特徴を考慮した分析支援ツールを提案する．RDF は分散した情報の結合や，構造化されていない情報の扱いが容易という特徴があり，活用事例が増えてきている．しかし，従来主流の関係データベースとは異なる特徴を有するため，専用の分析支援ツールが必要と考える．本稿では，構造化された部分データ空間の抽出，接続性の高いリソースの発見， ログデータを対象とした時系列データの抽出を行うツールを提案し，テキストデータマイニング統合環境 TETDM を用いて実装したプロトタイプを示す．

1. はじめに

本稿では RDF の特徴を考慮したデータ分析支援ツールを提案し，テキストデータマイニング統合開発環境 TETDM[1][4][5]にて実装したプロトタイプを示す．

Resource Description Framework (RDF) とは W3C[2] で規格化されている情報規格であり，リソースに関する情報を主語・述語 (プロパティ) ・目的語の 3 要素 (トリプル) で表現する．トリプルはグラフ構造で表すことが可能であり，述語や目的語の追加・更新を容易に行える．また，分散した情報の結合や，構造化されていない情報の扱いが容易という特徴がある．RDF で記述されたデータは，SPARQL などの RDF クエリ言語により，検索・操作することが可能である．

近年，RDF の特徴を生かし，ガバメント分野や科学技術分野などでデータを RDF で記述し活用する事例が増えてきている[6][7]．それに伴い，RDF データベースを分析する必要性が増してきていると考える．しかし，RDF データベースは複数のデータベースを横断して操作・検索を行うことが可能であることや，スケーラビリティが高く複雑なグラフ構造をしていることなど，従来主流の関係データベースとは異なる性質を有するため，既存のデータ分析ツールをそのまま適用することは難しく，専用の分析ツールが必要であると考える．

一方，分散されているデータマイニングツールを統一的に扱うことを目的とした，テキストデータマイニングによる統合開発環境 TETDM が公開されている．データマイニングのシステムやツールは，研究者・開発者が特定の分析に用いるために独自に開発しているケースが多く，また開発したツールは公開されない事も多い．TETDM はそういったツールを TETDM のモジュールとして公開・配布し，再利用することを目的としており，これにより分析者はツールの開発から始めることなく，分析に専念することが可能となる．

本稿では，RDF の特徴を考慮したデータ分析支援ツールを 3 種類提案する．提案する分析支援ツールは，構造化された部分データ空間を抽出し， 共通の述語を持つ主語の抽出･テーブル作成を行うツール，接続性の高いリソースの発見のために，複数エンドポイント間の共通リソース抽出を行うツール，そしてログデータを対象とした時系列データ抽出である．これらを TETDM のモジュールとして実装し，実際の RDF データベースに適用した事例を示す．

* 連絡先 首都大学東京 システムデザイン研究科
〒191-0065 東京都日野市旭ヶ丘 6-6
E-mail: tashiro-koichi@ed.tmu.ac.jp

[1] http://tetdm.jp/pukiwiki/index.php
[2] http://www.w3.org

2. 関連研究

2.1 RDF の活用事例

表 1 に，RDF データを公開しているサイトの一例を示す．欧米諸国では，政府や公共機関の統計データを，RDF を利用した Linked Open Data (LOD) として公開しており，代表的なものとして英国の DATA.GOV.UK や米国の DATA.GOV がある．また，Wikipedia[3] を RDF により記述した DBpedia も代表的な LOD である. 科学技術関連ではライフサイエンスの分野において，タンパク質の知識ベースとして，単一データベースではあるが UniProt が公開されている．さらに日本でも同様に，福井県鯖江市や，横浜市の芸術関連施設を公開している Yokohama Art Spot[12]などが RDF を採用している．また，日本語版 Wikipedia を RDF 化した DBpedia Japanese も公開されている．また，ライフサイエンスの分野では，データベースの統合が重要であるとの考えから，ヒト遺伝子データベース H-InvDB が公開されている[13]．さらには，食生活の管理・分析を行なう Web サービスとして FoodLog があり，ライフログデータの記述に RDF が用いられているなど, 様々な分野のデータに対して RDF が利用されている．

表 1: RDF の活用事例.

サイト名	URL
DATA.GOV.UK	http://www.data.gov.uk/
DATA.GOV	http://www.data.gov/
DBpedia	http://wiki.dbpedia.org/About
UniProt	http://www.uniprot.org/
福井県鯖江市	http://www.city.sabae.fukui.jp/pageview.html?id=11552
Yokohama Art Spot	http://lod.ac/apps/yas/
DBpedia Japanese	http://ja.dbpedia.org/
H-InvDB	http://www.h-invitational.jp/hinv/ahg-db/index_ja.jsp
FoodLog	http://www.foo-log.co.jp/index.html

2.2 RDF データ分析

現在，RDF データを対象とした分析に関する研究は，データ構造の理解を目的としたスキーマの可視化と，データ分布の理解を目的としたグラフ検索支援の 2 つに大別できる．以下ではそれぞれについてまとめる．

2.2.1 データ構造の理解

Goyal らは, RDF で記述されたデータを読み込み，情報関係をグラフ形式で表す RDF Gravity[4]を公開している． RDF Gravity は RDF クエリ言語の 1 つである RDQL によりトリプルを検索し，検索結果のグラフを表示することができる．Deligiannidis ら[14]は，データ構造とデータの理解を目的とし，RDF データ検索と可視化を行う Paged Graph Visualization (PGV) を提案している．通常，グラフ可視化ツールは，グラフ全体の可視化を行った後で，無関係なデータを除外し，目的とするデータの探索・可視化を行うものが多い．これに対し，PGV では， 小さなグラフから段階的に大規模な RDF オントロジ関連のデータを探索し，可視化を行う．

2.2.2 データ分布の理解

後藤ら[3]は，メタデータを対象とした探索的検索行為を支援する DashSearch LD を提案している．探索的検索とは，探索目的を少しずつ明確化して知識を獲得する情報検索であり, 検索空間を推移しつつ，途中の絞り込み検索をするという行為を繰り返し行う．これにより，情報要求の具現化だけでなく，検索空間を理解することができるため，検索に最適なクエリの入力が可能になるとしている．飯塚ら[1]は，複数の RDF のデータをマージして得られた単一のデータセットから頻出部分グラフパターンを抽出することで，RDF データを対象としたグラフ検索に必要なクエリを自動生成する手法を提案している．これにより，データ構造の把握をすることなく，適切なクエリの選択が可能となるため，類似するデータや比較対象となるデータの検索が容易になるとしている．

2.3 テキストデータマイニング統合開発環境 TETDM

TETDM (Total Environment for Text Data Mining) [4][5]とはテキストデータマイニングのための統合開発環境であり，Java で構築されている．複数のデータマイニング技術を柔軟に組み合わせて使える環境を目標としており，複数の研究者・技術者が別々に開発したデータ分析ツールをそれぞれモジュールとして扱うことが可能となっている．ユーザはモジュールがセットされたパネルを任意の枚数並べて分析を行う．各パネルにおいて，処理ツールと可視化

[3] http://www.wikipedia.org/

[4] http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html

ツールをそれぞれ1対1で組み合わせてセットする．また，複数のパネルを連動して利用することも可能であり，様々な観点から分析を行うことが可能である．

徳永ら[10][11]はオープンソースの統計解析を行う開発実行環境である R 言語[5] や，ニュージーランドのワイカト大学にて開発・公開されているデータマイニングツールである Weka[6] を，TETDM のモジュールとしてシステム化し，組み合わせて使うことを提案しており，TETDM と既存のデータマイニングツールの融合も進められている．

TETDM を用いたツールの開発事例として，高間ら[8]は，専門用語の候補を抽出し，専門用語辞書を作成する作業を支援するツールを提案している．また，梶並[2]は情報系の専門教育における TETDM の有効性を指摘し，実際に大学の講義において，受講者がモジュール開発を行った事例を報告している．

この様に，幅広い用途で TETDM は利用されており，今後も適用範囲がさらに広がることが期待されている．

3. 提案する分析支援ツール

3. 1 共通の述語を持つ主語の抽出・テーブル作成

従来のデータ分析やデータマイニングツールでは，表形式に構造化されたデータを対象とする場合が多い．これに対し，RDF データベースには，多くの述語 (プロパティ) が格納されており，その使用頻度は述語ごとに大きく異なることが一般的である．一般的な関係データベースにおいても欠損値が存在することはあるが，RDF データベースにおいて全ての述語を属性として表形式に構造化した場合，スパース度が非常に高くなる可能性が高い．従って，RDF データを分析するためにデータの構造化を行う場合，密度がある程度高い部分空間の抽出を行う必要があると考える．

本稿では，図 1 に示すように，最大公約数的に共通の述語を持つ主語の抽出を行い，行を主語，列を述語としたテーブルの出力を行うツールを提案する．これにより，データ分析に必要となる部分空間の抽出が可能になり、同時にどういった述語が多く使用されているかが把握できるので，データ分布の理解が可能となる．以下，部分空間抽出の手順を図 2 に示す具体例により説明する．始めに述語を抽出し，各述語を持つ主語の件数を求める．次に最大件数の述語とそれ以外の各述語において AND 検索を行ない，両述語を持つ主語の件数を求める．図 2 の例では，最大件数を持つ述語 B と他の述語それぞれについて AND 検索を行う．以降，最大件数をとる述語を AND 検索に追加していく．図 2 では，A が追加されている．AND 検索によるヒット数が全て 0 になるか，追加する述語がない場合，繰り返しを終了する．

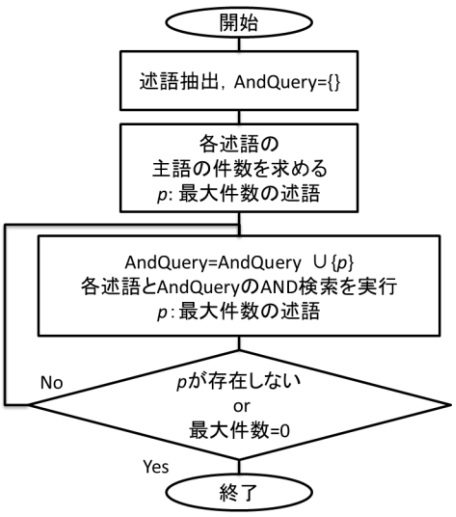

図 1: フローチャート．

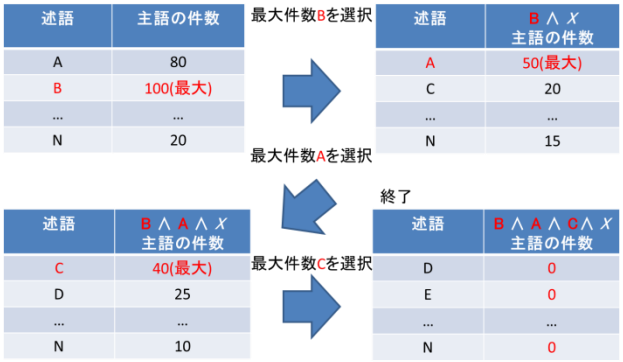

図 2: フローの具体例．

3. 2 複数エンドポイント間の共通リソース抽出

現在，LOD に代表されるように大量のデータが公開されており，それらのデータベースを横断して検索を行うことが，LOD 活用における一つの利点として挙げられる．Tim Berners-Lee は 2010 年の TED

[5] http://www.r-project.org/index.html
[6] http://www.cs.waikato.ac.nz/ml/weka/

University[9]にて，公開されている複数のデータを繋ぎ合わせることによって新たなデータの見方が可能になり，単体のデータからは得られない新たな結果が得られると述べている．しかし，複数のエンドポイント間のデータから意味のある結果，すなわち特徴的なリンク関係を発見するには，あらかじめデータ構造の把握が必要であり，単一のデータベースを対象とする場合と比較して，分析に手間がかかると考える．

そこで本稿では，接続性の高いリソース，すなわち 2 つの SPARQL エンドポイント間で共通するリソースを抽出し，そのリソースが持つ述語を把握するツールを提案する．これにより，複数データベースを横断した検索の手がかりを得ることが可能となり，データのリンク関係、すなわちデータ構造の理解につながることが期待できる．図 3 に示すように，2 つの SPARQL エンドポイントにおいて双方に出現しているリソースを共通リソースとして抽出する．また，各エンドポイントにおいて，共通リソースの主語あるいは目的語となる述語も同時に抽出する．

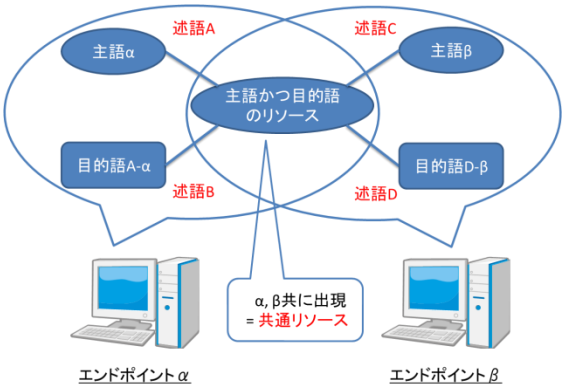

図 3: 複数エンドポイント間の共通リソース抽出．

3. 3 時間情報に基づくデータ分析支援

2.1 節で述べたように，RDF で記述されたデータには統計データやログデータも多くあり，それらには時系列データを扱っているものも多いため，時系列分析を行うことも想定される．あらかじめ分析したい時系列データが決まっていれば，そのデータのみ分析すればよいが，実際は様々な時系列データを探索的に分析したい場合も多いと考える．この場合には，時系列データの観点からデータ構造を理解する支援が有効であると考える．

そこで本稿では，時系列データを抽出し，ヒストグラムとして可視化するツールを提案する．時間情報が付与されている目的語を抽出し，分析対象とする．図 4 に示すように，目的語に時間情報をもつ述語を選択し，その述語の主語に対して他の述語，目的語を抽出する．図 5 に示すように抽出した述語と目的語から分析したい組み合わせを選択することにより，その目的語を横軸，時間の階級を縦軸としてヒストグラムを描画する．

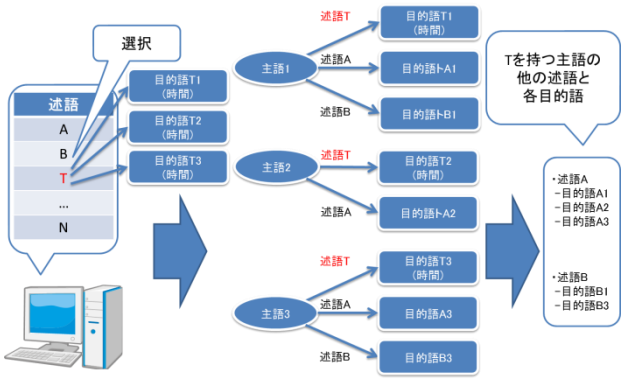

図 4: 時間情報に基づくデータ分析支援．

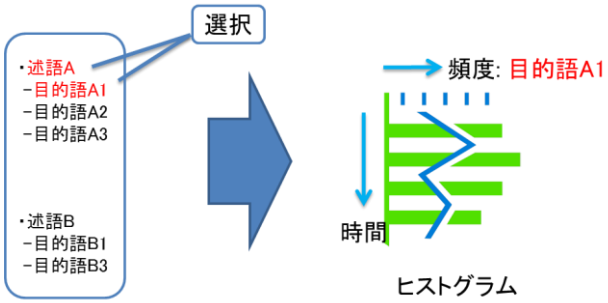

図 5: ヒストグラムにて可視化．

4. インタフェース

3 節で提案した 3 種類のツールについて，TETDM モジュールとして実装した．RDF の解析や SPARQL による問合せの処理に関しては，Apache Jena[7] を用いている．そのため，TETDM をコビルドする際に Jena の jar ファイルを含める様にしている．

図 6 に，3.1 節で提案した共通の述語を持つ主語の抽出・テーブル作成ツールのスクリーンショットを示す．このモジュールは，3.1 節のフローを行う処理ツールと，Type の表示・選択およびテーブルの表示を行う 2 種類の可視化ツールから構成される．モジュールの利用手順は以下の通りである．

1-1: SPARQL エンドポイントを入力
1-2: Type の表示及び選択
1-3: 検索する述語数の制限
2: テーブルを表示

[7] http://jena.apache.org/

3: テーブルをファイルに出力

ここで，ステップ 1-2，1-3 は任意であり，指定しない場合は全ての Type，述語を検索する．

図 7 は, ステップ 1 にて米国政府[8]の SPARQL エンドポイント を指定し，1-2 で述語 type の目的語として組織を表す FOAF の語彙である http://.../Organization を選択した場合の出力の一部である．この目的語を持つ主語は 176 件存在し，その中から主語 68 件, 述語 9 件のテーブルが抽出されている．

図 6: 共通の述語を持つ主語の抽出･テーブル作成ツールのスクリーンショット．

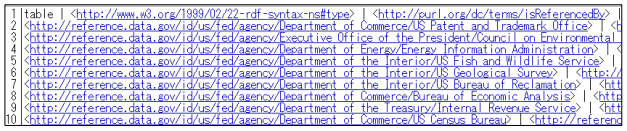

図 7: テーブルの出力例．

図 8 に, 3.2 節で提案した複数エンドポイント間の共通リソース抽出ツールのスクリーンショットを示す．このモジュールは，共通リソースとその共通リソースの述語を抽出する処理ツールと，その結果を表示する可視化ツールから構成される．モジュールの利用手順は以下の通りである．

1: 2 つの SPARQL エンドポイントを入力

2: 共通リソースを目的語に持つ述語 (A) ・共通リソース (B) ・共通リソースを主語に持つ述語 (C) の表示

図 9 は，ステップ 1 にて DBpedia Japanese [9]と DBpedia [10]の SPARQL エンドポイントを指定した場合の結果の一部である．http://dbpedia.org/ontology/Person が共通リソースの 1 つとして抽出されている．このリソースが関係する主語は両エンドポイントで共通であるが，目的語は異なっていることがわかる．

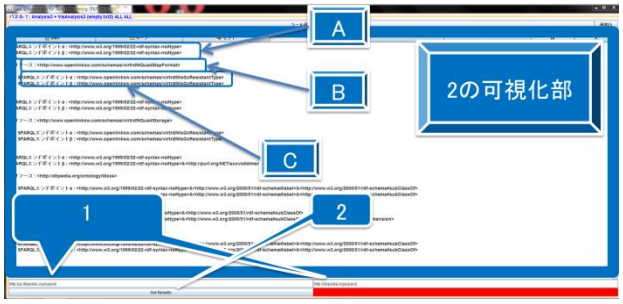

図 8: 複数エンドポイント間の共通リソース抽出ツールのスクリーンショット．

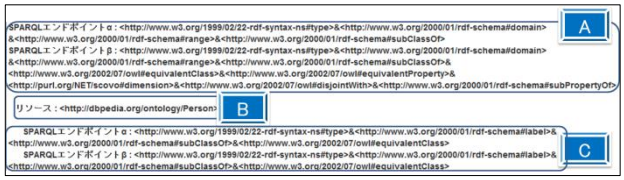

図 9: 共通リソースの抽出例．

図 10 に，3.3 節で提案した時間情報に基づくデータ分析支援ツールのスクリーンショットを示す．このモジュールは, 3.3 節で述べたヒストグラム生成を行う処理ツールと，述語・目的語の選択及びヒストグラムの描画をそれぞれ行う 2 種類の可視化ツールから構成される．現在では時間情報として XML Scheme の日付データ型である xsd:date，すなわち [yyyy-mm-dd]の形式の時間情報を持つデータに対して 5 つの階級幅において年の粒度での分析を可能としている．モジュールの利用手順は以下の通りである．

1-1: SPARQL エンドポイントを入力

1-2: 検索する述語数の制限

2: 述語の表示

3-1: 述語の選択

3-2: 選択した述語と共通の主語を持つ述語とその目的語の表示

4-1: 検索対象の述語の選択

4-2: 検索対象の目的語の選択

4-3: 年・階級間隔の入力

5: 選択した目的語の階級間隔内頻度の表示

ここで，ステップ 1-2 は任意であり，指定しない場合は全ての述語を検索する．

図 10 の左部は， ステップ 1-1 にて DBpedia Japanese，ステップ 3-1 で http://dbpedia.org/ontology/deathDate を選択し，ステップ 4-1 で http://dbpedia.org/ontology/deathPlace，ステップ 4-2 で http://ja.dbpedia.org/resource/東京都，ステップ 4-3 で年を 2000，階級間

[8] http://services.data.gov/sparql
[9] http://ja.dbpedia.org/sparql
[10] http://dbpedia.org/sparql

隔を 2 としそれぞれ指定した場合の結果である．このように時系列データの分析を行うことで，データ分析支援を行うことが可能である．

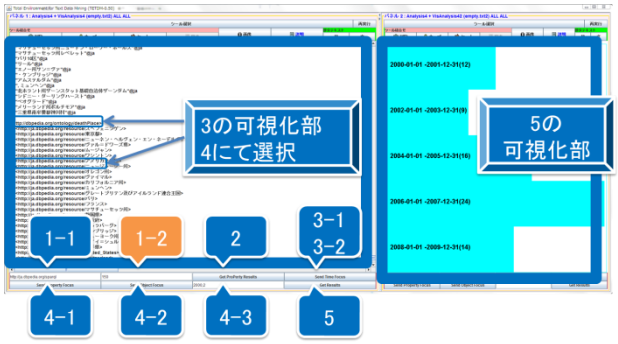

図 10: 時間情報に基づくデータ分析支援ツールのスクリーンショット．

5. まとめ

本稿では RDF の特徴を考慮した分析支援ツールを 3 種類提案し，TETDM を用いて実装したプロトタイプを示した．今後は各ツールの機能・インタフェースの向上のために，ユーザが指定可能な検索条件等の追加や，可視化表現の改善を行う予定である．3.1 節のツールに関しては，ユーザがテーブルの行・列の指定を可能にすることや，許容可能な欠損値の指定を可能とする予定である．3.2 節のツールに関しては，2 つ以上の SPARQL エンドポイントを指定可能にすることや，データ構造をよりわかりやすく提示することを考えている．3.2 節のツールに関しては，粒度としてより詳細な時間情報を扱えるように拡張する事を考えている．また，上記以外の新たなツールの開発や，開発したツールの有用性について評価・検討を行うことも予定している．特に様々なデータマイニングツールを統一的に扱うことを目的とした TETDM を用いて開発している意義として，提案・構築したツールを公開することで，フィードバックとしてより多くの意見を取り入れて反映させていくことも重要と考える．

参考文献

[1] 飯塚京士，佐藤宏之，イコプラムディオノ，村山隆彦: RDF データを対象としたグラフ検索におけるクエリ生成方式の検討，第 11 回セマンティックウェブとオントロジー研究会, SIG-SWO-A502-08, 2005

[2] 梶並知記: TETDM を利用した情報系専門教育の実践例，第 27 回人工知能学会全国大会，3B3-NFC-01a-5, 2013

[3] 後藤孝行，濱崎雅， 武田英明: DashSearch LD: 探索的検索の Linked Data への適用，第 26 回人工知能学会全国大会, 3C1-OS-13a-3, 2012

[4] 砂山渉，高間康史，ダヌシカボレラ，西原陽子， 徳永秀和，串間宗男，松下光範: テキストデータマイニングのための統合環境，第 25 回人工知能学会全国大会, 1B2-NFC3-10, 2011

[5] 砂山渉，高間康史，ダヌシカボレガラ，西原陽子，徳永秀和，串間宗男，松下光範: テキストデータマイニングのための統合環境 TETDM の開発，人工知能学会論文誌, Vol. 28, No. 1, pp. 1-12, 2013

[6] 総務省: 総務省におけるオープンデータに係る実証実験，http://www.opendata.gr.jp/committee/docs/20130122_4_rikatu.pdf，2013/09/14 現在

[7] 総務省: 総務省におけるオープンデータに関する技術の検討状況について，http://www.kantei.go.jp/jp/singi/it2/densi/wg/dai1/siryou8.pdf，2013/09/14 現在

[8] 高間康史，阿部美里: テキストデータマイニング統合環境を利用した看護記録からの専門用語辞書作成支援ツールの提案，第 27 回人工知能学会全国大会，3B3-NFC-01b-1, 2013

[9] Tim Berners-Lee: The year open data went worldwide, http://www.ted.com/talks/lang/en/tim_berners_lee_the_year_open_data_went_worldwide.html, 2013/09/14 現在

[10] 徳永秀和，杉村拓哉: R と Weka を活用した TETDM ツールの開発，人工知能学会情報編纂研究会第 6 回, TETDM-01-SIG-IC-06-07, 2011

[11] 徳永秀和: R によるテキストマイニング用 TETDM モジュール開発，第 27 回人工知能学会全国大会，3B3-NFC-01b-2, 2013

[12] 松村冬子，小林巌生，嘉村哲郎，加藤文彦，高橋徹，上田洋，大向一輝，武田英明: Linked Open Data による博物館情報および地域情報の連携活用,じんもんこん 2011 論文集, Vol. 2011, pp. 403-408, 2011

[13] 村上勝彦， 山崎千里， 今西規: ヒト遺伝子データベース H-InvDB の RDF 化と Endpoint の公開，第 27 回人工知能学会全国大会, 1N5-OS-10c-3, 2013

[14] Leonidas Deligiannidis, Krys j. Kochut, Amit P Sheth: RDF Data Exploration and Visualization, CIMS `07 Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience, pp. 39-46, 2007