様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １１ 日現在

機関番号：８２４０１
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０３２９８

研究課題名（和文）老化現象の解明に資する、オープンデータを体系的に利用した知識推論基盤の構築

研究課題名（英文）Development of reasoning technology of open data to contribute aging studies

研究代表者
桝屋　啓志（Masuya, Hiroshi）

国立研究開発法人理化学研究所・バイオリソース研究センター・室長

研究者番号：４０３２１８１４

交付決定額（研究期間全体）：（直接経費）　　13,600,000円

研究成果の概要（和文）：探索対象として化学物質オントロジーCHEBI、生命科学キーワードMESH、ヒト遺伝子リストHGNCを選定した。加えて、加齢を生理変化病的老化に分けて記述した老化オントロジーを作成した。
さらに、LODとテキストマイニングを融合した推論検索機能を開発した。文献共起データに基づき、キーワードと、キーワードを含むデータ（事象）との関係性を統計的に処理し（直接的関係層）、さらにデータ同士の関係を処理する（間接的処理層）をサーバー1台で実用速度で動作できるようにし、前述のオントロジーに基づいて作成された複雑なクエリを実行し、関連度情報をRDF等の形式で出力できるようにした。

研究成果の学術的意義や社会的意義
本研究の成果は、生命科学分野では全く新しいアプローチのプレ・スクリーニング技術のヒントとなると考えられる。従来、医薬品におけるリード化合物など、生理活性を持つ化合物選定の選定には、分子シミュレーションが用いられてきた。これに対して、本研究ではグローバルな知識ベースを参照して関連知識を抽出するという従来とは全く異なる新たなアプローチに取り組んだ。この方法の利点は、1) 検索条件を「特定の分子形状」ではなく、「アンチエイジング」等の最終目的に近い課題で設定できる。2) 広範囲かつ多様な情報を探索源とした知識抽出を行うことから汎用性が高いことが挙げられる。

研究成果の概要（英文）：We selected the chemical substance ontology CHEBI, the life science keyword MESH, and the human gene list HGNC as search targets. In addition, we created an aging ontology that describes aging separately for physiological changes and pathological aging.
In addition, we have developed an inference search function that combines LOD and text mining. Based on the literature co-occurrence data, the server statistically processes the relationship between the keyword and the data (event) containing the keyword (direct relationship layer), and further processes the relationship between the data (indirect processing layer). It is possible to operate at a practical speed with one unit, execute a complicated query created based on the above-mentioned ontology, and output relevance information in a format such as RDF.

研究分野：セマンティックウェブ

キーワード：オントロジー　知識探索

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、情報のオープン化やビッグデータ化が進み、多種多様なオープンデータの利活用を高めるための技術ニーズが高まっている。特に、肥大化するデータ量に埋もれて利用されない知識も増大しており、ウェブ上に存在する既存知識の大規模抽出、対象とする事象に関する知識動向の把握など、新規発見につながる推論や気づきを与える仕組みが求められている。

生命科学では他分野からの知見導入により研究のブレイクスルーがもたらされてきたことから、データ蓄積およびオープン化、データの相互結合による Linked Open Data (LOD)化が大きく進み、データ利活用法の開発の場としても注目されている。しかしながら、現在の情報統合技術では、蓄積されたデータからトランスレートされる新たな知識発見や気づきの提示は期待されたほどには進んでいない。

その理由の 1 つとして、現状の生命科学 LOD の内容の傾向として、実験結果や、「分野内」での知見集積等の確定知識が多いことがあげられる。一方、分野横断的な知識導入につながる情報は文献という異なる種類（モダリティ）のデータ内に豊富に記載されているのに、シームレスにこの情報にアクセスする手段が無い。また、「老化」等の生命の複合的な現象に関して、全体と部分から成る複雑な概念構造、及び各部分要素の依存するコンテキストを正しく扱えていないために、求められているような知識の抽出に結びついていないことも問題である。この問題の解決のためには、下記のように、LOD に基づく情報統合技術、テキストマイニング技術、オントロジーによる知識モデリング技術の融合が必要である。

- LOD とテキスト等異なるデータを、それぞれの利点を生かしながら横断的に扱う（LOD に基づく知識推論とテキストマイニング技術の融合）
- 複合的な知識について、コンテキスト等を区別するモデリングを正しく行った上で、そのモデルに準じた高度なクエリを行う技術（知識モデリングと検索技術の融合）
- 上記 2 つを組み合わせた上で、直接的関係性から導かれる既知の知見と、間接的関連性から擬似的に導かれる「未知」の知見を抽出し、検索結果をわかりやすく提示する。

この問題は生命科学だけでなく、多種混合のオープンデータを効率的に探索し利活用するための情報学的な課題でもある。

２．研究の目的

我々は、上記課題の解決に向けて、生命科学の老化研究への貢献を具体的例題として取り組む。すなわち、世界の研究機関からウェブ上で公開される知識を参照し、老化への関与が期待される化合物、遺伝子、バイオリソース等の選別を可能とする推論検索技術を開発する。老化という社会的課題に取り組むことで、異なる技術の融合により、多種混合のオープンデータを効率的に探索し利活用するための汎用的な情報技術開発を行い、インターネット上に蓄積されるデータの利活用に新たな局面を拓く

３．研究の方法

上記の目的に向けて、本研究ではテキストマイニング、知識推論、LOD 関連技術、オントロジーによる知識モデリングにおける国内トップレベル研究者による研究グループを組織し、各技術を組み合わせることで、LOD と文献を網羅的に探索し、老化に関連する事象を提示する実用的な推論技術を確立ために、、下記の 3 つの課題に分けて取り組んだ。

（１）探索対象データの整備

探索対象として本システムに含めるデータ(LOD)を、生命科学データベースである NBDC RDFPortal のデータや Integbio データベースカタログ、BioSharing.org などを対象に選別を行う。対象は、遺伝子、タンパク質、脂質、代謝物、化合物、バイオリソースの公共データベースとする。選別は各分野のエキスパートである桝屋と山田が行い、各データは、メタデータ記述のデファクトスタンダートである Resource Description Framework (RDF)形式に変換する。

（２）老化現象を記述するオントロジーの作成

老化現象を構成する多様な事象を、is-a （分類/継承）及び part-of （部分）の各関係性で整理し体系的にまとめたオントロジーを作成する。桝屋および古崎が先行研究で行ってきたオントロジー研究の成果を生かし、オントロジーの理論的な側面と，検索・ランキングというプラクティカルな側面の両面から優れた老化オントロジーを構築した。

（３）多様な大規模データを関係リンクを経由して跨ぎながら、キーワードとデータ（事象）間の関連スコアを高速に計算する推論検索機能の実現

先行研究において開発した推論システム PosMed（後述）のアルゴリズムを改良することで、LOD

とテキストマイニング由来データを融合した推論検索機能を実現する。PosMed では、文献共起データに基づき、キーワードと、キーワードを含むデータ（事象）との関係性を統計的に処理し（直接的関係層）、さらにデータ同士の関係を処理する（間接的処理層）を設けることで、最終的に統計的に評価されたランキングとして示すことができる。しかし、PosMed では間接的処理層が高々１段の関連リンクのみのため、収集される LOD を取り込みつつ、LOD の複数の関連リンクを扱える新たなシステムとする。さらにこの処理をサーバー1 台で実用速度で動作できるように軽量化等の改良を試みた。また推論アルゴリズムを、直接的関係層と間接的関係層に分けて検索結果を並列して示せるように機能改変する。これを擬似的に既知と未知の知見とし、それぞれの統計的評価方法を確立した。

４．研究成果

探索対象データの整備：探索対象として本システムに含めるデータ(LOD)を、生命科学データベースである NBDC RDF Portal のデータや Integbio データベースカタログ、BioSharing.org などを対象に選別を行った。その結果、化学物質の網羅的オントロジーである CHEBI、生命科学キーワードの語彙集である MESH、ヒト遺伝子のリストである HGNC を利用することを決定した。また、ヒト、マウス、ラットのタンパク、遺伝子名のデータの整備を行った。

老化オントロジー作成：老化現象を構成する多様な事象を、is-a （分類/継承）及び part-of（部分）の各関係性で整理し体系的にまとめたオントロジーを作成することを目的として、老化プロセスに従って機能する遺伝子、疾患等をリストアップするとともに、Interlinking Ontology forBiological Concepts (IOBC) を用いて抽出する作業をおこなった。これらをまとめることで、加齢による生理変化と病的因子によって進行する病的老化に分けて、約 350 の概念を記述した老化オントロジーを作成した。老化において観察される機能の低下、病名等を、症状、原因と関連づけて分類した。本オントロジーは、後述の検索システムのクエリに利用した。

推論アルゴリズムの開発：先行研究において開発した推論システム PosMed のアルゴリズムを改良することで、LOD とテキストマイニング由来データを融合した推論検索機能を実現することを目的として、文献共起データに基づき、キーワードと、キーワードを含むデータ（事象）との関係性を統計的に処理し（直接的関係層）、さらにデータ同士の関係を処理する（間接的処理層）をサーバー1 台で実用速度で動作できるように軽量化を行なった。これを用いて予備的な検証を行った。二次代謝産物データベースの 49,983 件の化合物リストと、Medline2018 最新版の 2,456 万件の文献データを用いて、老化キーワードと、直接ではなく、遺伝子や疾患を介して間接的に関連するデータを抽出し、関連度によりランキングした。それにより抽出された上位 10 個の化合物について Pubmed を調査したところ、老化と直接の関連性は見つからず、新規老化関連化合物の候補となる可能性が示唆された。また、操作に関する GUI の整理を行い、様々な要素の関連ををわかりやすく提示できるようにした。

さらに、この推論検索機能の拡張として、直接的関係層と間接的処理層を用いた推論検索機能について以下の 4 つの目的で拡張を行った。

（１）約 2000 万件の生命科学系の論文セットに対してキーワード検索し、ヒットした文献数を得る。

（２）RDF 形式で記述されたデータセットに対して、（１）の機能を使ってデータと論文との関連づけを文献数に基づいた統計的手法により行う。

（３）（１） と（２）の機能を組み合わせ、2 つのデータセットについてデータ同士の関連付けを文献数に基づいた統計的手法により行う。

（４）複数のデータセットを Aging Miner に組み込んだ時、（１）（２）（３）のそれぞれの機能を組み合わせ、 (A) 文献→データセット、(B) 文献→データセット→データセット、(C)文献→データセット→データセット→データセットの経路でデータを検索できる様にする。これらを実現するために、データの構築は MySQL 等の RDB を用いて行ない、検索システムは、Lucene 等のテキスト検索エンジンと、Virtuoso 等の RDF データベースのハイブリッドとする設計を行った。RDB のテーブルは各々のステップで統合された静的なテーブルを用いた。テキスト検索エンジン＋RDF の環境は１ つのプロセスとして構築し、途中復帰が可能な様に各々のステップで情報を RDB に保存するようにした。

さらに、これらをまとめて１ つのツールとするプロトタイプを作成した。本ツールは、オントロジーに基づいて作成された複雑なクエリをバッチ検索として実行し、膨大な文献情報に対して、検索キーワードに対する、遺伝子、バイオリソース、疾患、代謝産物、薬剤等との関連度情報を Resource Description Framework (RDF)あるいはそれに準ずる形式で出力し、その結果を老化研究に資する Open Linked Data として公開できるように設計されており、老化関連データを検索するシステム Aging Miner として公開予定である。

検索ジョブには、検索式、ターゲット検索経路を最大 500 組投入でき、検索経路は(A) 文献→データセット、(B) 文献→データセット→データセット、(C)文献→データセット→データセット→データセットの経路のいずれかを選ぶことが可能である。検索結果は JSON, JSON-LD, および ttl 形式で出力される。これを用いることで、老化に紐づく遺伝子、バイオリソース、疾患、代謝産物、薬剤等を、根拠となる論文との関連性も含めた結果を RDF として出力し、その知識

を記述した RDF を様々なシステムで運用可能となる。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Tanaka Nobuhiko、Masuya Hiroshi	４．巻 10
２．論文標題 An atlas of evidence-based phenotypic associations across the mouse phenome	５．発行年 2020年
３．雑誌名 Scientific Reports	６．最初と最後の頁 3957
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-020-60891-w	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Haselimashhadi Hamed et al.	４．巻 1
２．論文標題 Soft windowing application to improve analysis of high-throughput phenotyping data	５．発行年 2019年
３．雑誌名 Bioinformatics	６．最初と最後の頁 1492-1500
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btz744	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Furuse Tamio、Mizuma Hiroshi、Hirose Yuuki、Kushida Tomoko、Yamada Ikuko、Miura Ikuo、Masuya Hiroshi、Funato Hiromasa、Yanagisawa Masashi、Onoe Hirotaka、Wakana Shigeharu	４．巻 12
２．論文標題 A new mouse model of GLUT1 deficiency syndrome exhibits abnormal sleep-wake patterns and alterations of glucose kinetics in the brain	５．発行年 2019年
３．雑誌名 Disease Models & Mechanisms	６．最初と最後の頁 38828
掲載論文のDOI（デジタルオブジェクト識別子） 10.1242/dmm.038828	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tanaka Nobuhiko、Masuya Hiroshi	４．巻 2018
２．論文標題 Mouse phenome as biological resource	５．発行年 2018年
３．雑誌名 Impact	６．最初と最後の頁 93～95
掲載論文のDOI（デジタルオブジェクト識別子） 10.21820/23987073.2018.12.93	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計14件（うち招待講演　3件／うち国際学会　5件）

１．発表者名
櫛田達矢、岩瀬秀、湯原直美、栗原恵子、並木由理、臼田大輝、高田豊行、田中信彦、鈴木健大、桝屋啓志

２．発表標題
バイオリソースの利活用向上を目指した理研BRCホームページのコンテンツ整備

３．学会等名
トーゴーの日シンポジウム2020

４．発表年
2020年

１．発表者名
臼田大輝、櫛田達矢、小林紀郎、桝屋啓志

２．発表標題
セマンティックWeb技術を用いた、バイオリソースカタログシステムの開発と運用

３．学会等名
2020年度人工知能学会全国大会（第34回）

４．発表年
2020年

１．発表者名
臼田大輝、櫛田達矢、小林紀郎、桝屋啓志

２．発表標題
理研BRCのバイオリソース種横断的に検索可能なカタログシステムの開発

３．学会等名
トーゴーの日シンポジウム2020

４．発表年
2020年

１．発表者名
臼田大輝、櫛田達矢、小林紀郎、桝屋啓志

２．発表標題
複数種類のバイオリソースを横断的に検索するカタログシステムの開発

３．学会等名
第67回日本実験動物学会総会

４．発表年
2020年

１．発表者名
桝屋啓志

２．発表標題
情報と一体化した高付加価値リソース創出に向けて

３．学会等名
第42回日本分子生物学会年会 ナショナルバイオリソースプロジェクト公開シンポジウム（招待講演）

４．発表年
2019年

１．発表者名
Hiroshi Masuya, Daiki Usuda, Naomi Yuhara, Keiko Kurihara, Yuri Namiki, Shigeru Iwase, Kenta Suzuk and Nobuhiko Tanaka

２．発表標題
Homepage of RIKEN BioResource Research Center

３．学会等名
33rd International Mammalian Genome Conference（国際学会）

４．発表年
2019年

１．発表者名
桝屋啓志、岩瀬秀、田中信彦、鈴木健大、湯原直美、臼田大輝、栗原恵子、並木由理、佐藤道比古、小幡裕一

２．発表標題
バイオリソース研究センターにおける情報統合

３．学会等名
第32回モロシヌス研究会

４．発表年
2019年

１．発表者名
Hiroshi Masuya, Daiki Usuda, Naomi Yuhara, Keiko Kurihara, Yuri Namiki, Shigeru Iwase, Kenta Suzuk and Nobuhiko Tanaka

２．発表標題
Data integration for bioresource

３．学会等名
INFRAFRONTIER/IMPC Conference 2019（国際学会）

４．発表年
2019年

１．発表者名
桝屋啓志、岩瀬秀、田中信彦、鈴木健大、湯原直美、臼田大輝、栗原恵子、並木由理、佐藤道比古、小幡裕一

２．発表標題
理研バイオリソース研究センターにおける情報整備

３．学会等名
第66回日本実験動物学会総会

４．発表年
2019年

１．発表者名
桝屋啓志，古崎晃司，溝口理一郎

２．発表標題
視点依存オープンデータの大規模作成支援ツール

３．学会等名
人口知能学会全国大会（第32回）

４．発表年
2018年

１．発表者名
山口敦子，小林紀郎，桝屋啓志，山本泰智，古崎晃司

２．発表標題
LOD Surfer API: クラス間関係に基づく LOD 探索のためのウェブ API

３．学会等名
2018年度人口知能学会全国大会（第32回）

４．発表年
2018年

１．発表者名
Masuya H

２．発表標題
Data integration in RIKEN BioResource Research Center

３．学会等名
The 10th ANRRC International Meeting（国際学会）

４．発表年
2018年

１．発表者名
Masuya H

２．発表標題
RDF-based data integration of mouse phenotype

３．学会等名
6th INCF Japan Node International Workshop（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Masuya H

２．発表標題
Data integration of mouse phenotype

３．学会等名
AMMRA & AMPC Meeting Workshop（招待講演）（国際学会）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	古崎　晃司 (Kozaki Kouji) (00362624)	大阪電気通信大学・情報通信工学部・教授 (34412)	
研究分担者	小林　紀郎 (Norio Kobayashi) (20415160)	国立研究開発法人理化学研究所・情報システム本部・ユニットリーダー (82401)	
研究分担者	山田　一作（小山内一作） (Issaku Yamada) (50370185)	公益財団法人野口研究所・研究部・研究員 (72690)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関