Wikidataを中心としたデータ統合に向けた専門知識の包含関係の分析

6U-05

峯 拓也 † 古崎 晃司 ‡
大阪電気通信大学 † 大阪電気通信大学 ‡

1 研究背景

知識グラフとは知識の関係をグラフ構造で表したデータベースであり、知識型アプローチによるAIシステムを開発するため重要とされる。これはAIシステムが対象とするタスクによっては人間が持っている知識を適切に表現することが必要であり、その用途に知識グラフが適しているためである。知識グラフの特徴は、知識の関係性が明確に表現されている点にあり、その関係性を利用して、複数の知識グラフのデータを統合することにより情報を強化できる。

先行研究 [1, 2, 3] などでは、知識グラフの1つであるWikidataを使用し、Wikidataをデータ統合のハブとして活用できるかについて、統計的に分析をしている。しかし、分野を限定せずWikidata全体として分析を行っているため、分野による違いを分析する必要がある。そこで本研究では、Wikidata内でもデータ数が豊富なライフサイエンス分野の中から、社会的な問題などを考えると重要であると考えられる疾患を例として、Wikidataを中心としたデータ統合について分析する。

2 目的

本研究の目的は、「Wikidataを中心とした専門知識のデータ統合」であり、本稿ではWikidataにおける専門知識に関するデータが持つ外部データベースへのリンク情報を用いることでどのようなデータ統合を行うことができるかを考える。対象を限定しない汎用的な知識グラフであるが、各データから様々な専門領域の外部データベースにリンクされており、外部の様々なデータベースをつなぐハブの役割を果たしている。その為、Wikidataを中心とし、データを統合することによりデータの専門知識の収集が行えると考えられる。

その際に、Wikidataに定義されている専門知識と、外部データベースに定義されている専門知識のカバー範囲の違いが統合されたデータのカバー範囲に影響するため、それらの包含関係を分析する。その分析結果を通してWikidataが専門知識のデータ統合にどの程度、用いることができるのかを考察する。

3 使用技術

3.1 LOD(Linked Open Data)

Linked Data(データ同士が相互にリンクされているデータ)として公開されたOpenData(誰でも自由に使えるよう公開されているデータ)でありRDFで構造化されているデータベースとなっており、また他のLODとの繋がりを持っている。Wikidataは代表的なLODの1つである。

3.2 Wikidata

Wikidata[4]とは、Wikipediaと同じように誰でも編集できる知識ベースで、Wikipediaのデータ版とも位置付けることができる。項目数は1億を超えている(2023年1月時点)オープンな知識ベースであり、SPARQLエンドポイントや各種検索ツールの提供がされている。Wikidataのデータ表現は、主語(subject)、述語(predicate)、目的語(object)の3要素の組み合わせで表されるRDF(Resource Description Framework)による構造化が用いられており、これらのグラフ構造の組み合わせにより検索を行える。各データは、Q…というIDを用いて表されるデータ(リソース)と、P…というIDを用いて表されるデータ間の関係(プロパティ)からなり、それぞれの省略表現としてして、wd:Q12136(病気)やwdt:P31(分類)といった表現が用いられる。

3.3 MeSH

MeSH(メッシュ)は、Medical Subject Headingsの頭文字であり、米国国立医学図書館(NLM)が定める生命科学用語集である。NLMが文献を管理する際、文献の内容を表す適切な用語を10〜15個程度文献に付与し、この用語により文献を検索・管理できるようにしている。MeSHは毎年改訂されており新しい概念や語句が追加・修正され、最新の生命科学に対応できるようにされている。MeSHは、定義語、副表題、補足用語の3つの基本事項と、それに加えて出版物の種類、場所についての用語を含んでいる。また、MeSHは階層構造になっており、下位にいくほど厳密な定義語となる。

3.4 DiseaseOntology

Disease Ontologyは、人間の病気に関するオントロジーでメリーランド大学医学部のゲノム科学研究所で運営されている。このプロジェクトは、生物医学リポジトリにアノテーションされたあらゆる疾患概念を、コミュニティのニーズに合わせて拡張可能なオントロジーフレームワークでカバーすることを目的としたオントロジーへのニーズに応えるものである。またDisease Ontologyは生命科学分野のオントロジーを共有するための学術的なコミュニティであるOBO (Open Biomedical Ontologies) Foundryにも登録されている。このオントロジーでIDとして利用されるDisease Ontology Identifiers (DOID)は、接頭辞DOID:の後に番号が付いたもので、例えばアルツハイマー病は識別子DOID:10652を持っている。また、DOはUniProtなどいくつかのリソースで相互参照されている。

4 包含関係の分析

4.1 分析手順

Wikidataと外部データベース(外部DB)における疾患について専門知識の包含関係を調べるための分析の手順を下記に示す。

(1) Wikidataから疾患の一覧(WD疾患)の取得
(2) (1)で得た一覧から、それぞれの対応する外部DBのIDを取得
(3) 外部DBから疾患の一覧(外部DB疾患)の取得
(4) (1)-(3)の包含関係を分析

抽出手順(1)では、Wikidataの疾患一覧の取得を行う。SPARQLクエリを用いて目的語を「病気(wd:Q12136)または、上位クラス(wdt:P279)が病気となるクラス」とし、述語が「(wdt:P31)」となる主語を持つグラフ構造(図1)を検索することで疾患の一覧を取得しその結果14451件のWikidata疾患が得られた。

図1 Wikidataから疾患一覧の抽出

手順(2)では、Wikidataから外部DBへのリンク情報を用いて、(1)で取得したWD疾患リンクされている外部DBでそれぞれに対応するIDを取得する対象とする外部DBを選定するにあたり、WD疾患から外部DBへのすべてのリンクを調べ全体の件数が多くRDF化されているデータを優先することとした。WD疾患からのリンクの件数が多い外部DBを調べた結果の一部を、図2に示す。

外部DBへのリンク	リンク先の外部ID	件数
wdt:P4229	ICD-10-CM	5079
wdt:P486	MeSH descriptor ID	4609
wdt:P699	Disease Ontology ID	4567
wdt:P7807	ICD-11 (foundation)	3043
wdt:P665	KEGG ID	1826

図2 WD疾患からリンク件数が多い外部DB

今回は、図2に示した外部DBのうち、RDF化されているデータベースであるMeSH、DiseaseOntology、の2つを分析対象とすることにした。これらの外部DBでのIDは、WD疾患からそれぞれの外部DBへのリンク先となっているIDを取得するSPARQLクエリを実行することで得られる。

手順(3)では、各外部BDの疾患一覧の取得を行う。各DB内でデータの分類が疾患となるものを検索するSPARQLクエリを用いて疾患一覧の取得を行う。

手順(4)では、手順(4) (1)-(3)で得られた一覧を用いて、包含関係を調べる。WD疾患と外部DBの疾患の対応は、手順(2),(3)で用いたWikidataから外部DBへのリンクによるIDの対応関係を用いて判断する。WD疾患と外部DB疾患の双方に含まれるもの、片方のみに含まれるものの一覧を取得し、各DBごとの違いを見る。包含関係を表にまとめると図3のようになり、外部DBからWikidataへのIDリンクがなかった場合疾患として良いのかなどを見る。

Analyzing the inclusion of expertise for data integration with a focus on Wikidata
† MineTakuya,Osaka Electro-Communication University
‡ KozakiKouji,Osaka Electro-Communication University

なお、手順 (2),(3) において、Wikidata の SPARQL クエリサービスを用い Wikidata での疾患一覧抽出し得た一覧から、それぞれの対応する外部 DB の ID を取得を行う部分は、外部 DB へのリンクの種類を変更するだけで行えるようにした。

Wikidata	外部DB
○	○
○	×
×	○

図 3　(1)-(4) により得られた結果包含関係を簡略化したもの

4.2　MeSH を利用した情報の抽出

手順 (2) では、図 1 のグラフ構造から得られる Wikidata の疾患一覧から MeSH へのリンクを表す wdt:P486(MeSH descriptor ID) の目的語（リンク先）となる MeSH の ID を調べる SPARQL クエリを作成し ID の一覧を取得した。グラフ構造で表すと図 4 のようになる。取得した MeSH の ID の総数は 10855 件であった。

図 4　Wikidata から MeSH の ID リンク取得

手順 (3) では、MeSH で疾患を表す分類である meshv：SCR_Disease を用いて SPARQL クエリの作成を行い疾患一覧を取得した。クエリの内容は、目的語を「meshv：SCR_Disease」とし、述語が「type」となる主語を検索（抽出）するものであり、グラフ構造で表すと図 5 のようになる。取得した全体の総数は 19420 件であった。

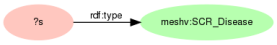

図 5　MeSH から疾患の一覧取得

4.3　DiseaseOntology を利用した情報の抽出

MeSH と同様に (2) の手順を用い、Wikidata の疾患一覧から DiseaseOntology へのリンクを表す wdt:P699(Disease Ontology ID) をグラフ構造で表すと図 6 のようになりの図 4 の P486 を変更することにより疾患を調べる SPARQL クエリを作成し ID の一覧を取得した。取得した DiseaseOntology の ID の総数は 12287 件であった。

図 6　DiseaseOntology からの疾患一覧取得

手順 (3) では、DiseaseOntology は疾患に特化したデータベースであるため、DiseaseOntology 内での疾患 ID を表す述語を使用し、疾患 ID を持つデータを疾患一覧として取得した。グラフ構造で表すと図 4 のようになる。取得した全体の総数は 13647 件であった。

図 7　DiseaseOntology からの疾患一覧取得

5　結果

得られた結果から各 DB との包含関係を調べた、Wikidata と MeSH について、Wikidata と DiseaseOntology について 2 つの結果を示す。

5.1　Wikidata、MeSH について

Wikidata と MeSH の包含関係を分析した結果は双方に存在している疾患の数は 3009 件、Wikidata のみに存在している疾患の数は 3805 件、MeSH のみに存在している疾患の数は 3501 件となった（図 8)。この結果より、MeSH 内での疾患（13647 件）のうち、22 %が Wikidata にカバーされていることが分かる。

一方、MeSH のみに存在しているについては、MeSH 内での疾患（13647 件）のうち、25 %が Wikidata にカバーされていることが分かる。専門知識では疾患とされているが Wikidata 上での分類が違う可能性があるためでが、Wikidata 上での定義内容を確認することにより疾患知識として扱うことができる可能性がある。続いて、Wikidata のみに存在している疾患について、MeSH 上での分類 (type) を調べた結果、3805 件の内 TopicalDescriptor が 3740 件、SCR_Chemical が 3 件、その他 62 件 (図 9) となっていた、TopicalDescriptor は、MeSH において他の用語を修飾するための語彙をあらわすため、これらは疾患を表す修飾子であると思われる。その為、これらの用語についても疾患として扱うことができると思われる。

Wikidata	MeSH	件数
○	○	3009
○	×	3805
×	○	3501

図 8　Wikidata と MeSH 関係

type	件数
<http://id.nlm.nih.gov/mesh/vocab#TopicalDescriptor>	3740
<http://id.nlm.nih.gov/mesh/vocab#SCR_Chemical>	3
その他	62

図 9　Wikidata のみに存在している疾患 type

5.2　Wikidata、DiseaseOntology について

Wikidata と DiseaseOntology の双方に存在している疾患の数として 10551 件、Wikidata のみ存在している疾患の数として 2 件、DiseaseOntology のみ存在している疾患の数として 3096 件という結果を得ることができた（図 10)。この結果より、DiseaseOntology での疾患 (12287 件) のうち、85 %が Wikidata にカバーされていることが分かる。

DiseaseOntology のみ存在している疾患については、MeSH と同様に、Wikidata 上での分類の確認することで、疾患知識として扱うことができる可能性がある。Wikidata のみ存在している疾患 2 件に関しては、1 つのデータについては DiseaseOntology 側にデータが存在しておらず 1 つののデータに関しては DiseaseOntology 上にデータは存在し、Wikidata にリンクされていいなかった。これは、Wikidata でのリンクを付与する際にマッチングがうまく取れなかった可能性がある。

Wikidata	DiseaseOntology	件数
○	○	10551
○	×	2
×	○	3096

図 10　Wikidata と Diseaseontology 関係

6　結論

本研究では、Wikidata を中心に、MeSH、DiseaseOntology の 3 つの知識グラフを用いて疾患に関する情報の抽出を行い包含関係を調べた。

Wikidata と MeSH では、双方に存在する疾患、Wikidata のみに存在している疾患の数 6510 件に関して、約半数は Wikidata 上で疾患とされている。これは統合に用いる際疾患としてよいと考える。また、Wikidata のみに存在している疾患に関して、TopicalDescriptor という疾患を表す修飾子を持つものに対しては同様に疾患としてよいと考える。

Wikidata と DiseaseOntology では、双方に存在している疾患は、Wikidata から得られた疾患数の約 8 割カバーできている。DiseaseOntology のみ存在している疾患は、専門知識から取得しているため疾患とみなすことができる。現在は、RDF 化されているデータベースから抽出しているがそれら以外の専門知識データベースからも API などを利用することにより同様の抽出方法が使用できると考え、関係を調べることにより統合を行えるようにしていく必要があると考えられる。

参考文献

[1] AnAnalysisofLinksinWikidata ArminHaller,AxelPolleres,DaniilDobriy,NicolasFerranti ,SergioJ.Rodr ´ guezMendez ´/ Australian National University,Vienna University of Economics and Business

[2] Semantic Data Integration for Knowledge Graph Construction at Query Time Diego Collarana; Mikhail Galkin; Ignacio Traverso-Ribón; Christoph Lange; Maria-Esther Vidal; Sören Auer 2017 IEEE 11th International Conference on Semantic Computing (ICSC)

[3] Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction Association for Computational Linguistics 3219–3232 Yi Luan Luheng He Mari Ostendorf Hannaneh Hajishirzi

[4] https://www.wikidata.org/wiki/Wikidata: