Original Paper

Leveraging Knowledge Graphs and Natural Language Processing for Automated Web Resource Labeling and Knowledge Mobilization in Neurodevelopmental Disorders: Development and Usability Study

Jeremy Costello[1], BEng; Manpreet Kaur[2], MEng; Marek Z Reformat[1,3], MSc (Hons), PhD; Francois V Bolduc[2,4,5,6], MD, PhD

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada
[2]Department of Pediatrics, University of Alberta, Edmonton, AB, Canada
[3]Information Technology Institute, University of Social Sciences, Łódź, Poland
[4]Department of Medical Genetics, University of Alberta, Edmonton, AB, Canada
[5]Women and Children Health Research Institute, University of Alberta, Edmonton, AB, Canada
[6]Neuroscience and Mental Health Research Institute, University of Alberta, Edmonton, AB, Canada

Corresponding Author:
Francois V Bolduc, MD, PhD
Department of Pediatrics
University of Alberta
3-020 Katz Building
11315 87 Avenue
Edmonton, AB
Canada
Phone: 1 780 492 9713
Email: fbolduc@ualberta.ca

Abstract

Background: Patients and families need to be provided with trusted information more than ever with the abundance of online information. Several organizations aim to build databases that can be searched based on the needs of target groups. One such group is individuals with neurodevelopmental disorders (NDDs) and their families. NDDs affect up to 18% of the population and have major social and economic impacts. The current limitations in communicating information for individuals with NDDs include the absence of shared terminology and the lack of efficient labeling processes for web resources. Because of these limitations, health professionals, support groups, and families are unable to share, combine, and access resources.

Objective: We aimed to develop a natural language–based pipeline to label resources by leveraging standard and free-text vocabularies obtained through text analysis, and then represent those resources as a weighted knowledge graph.

Methods: Using a combination of experts and service/organization databases, we created a data set of web resources for NDDs. Text from these websites was scraped and collected into a corpus of textual data on NDDs. This corpus was used to construct a knowledge graph suitable for use by both experts and nonexperts. Named entity recognition, topic modeling, document classification, and location detection were used to extract knowledge from the corpus.

Results: We developed a resource annotation pipeline using diverse natural language processing algorithms to annotate web resources and stored them in a structured knowledge graph. The graph contained 78,181 annotations obtained from the combination of standard terminologies and a free-text vocabulary obtained using topic modeling. An application of the constructed knowledge graph is a resource search interface using the ordered weighted averaging operator to rank resources based on a user query.

Conclusions: We developed an automated labeling pipeline for web resources on NDDs. This work showcases how artificial intelligence–based methods, such as natural language processing and knowledge graphs for information representation, can enhance knowledge extraction and mobilization, and could be used in other fields of medicine.

(J Med Internet Res 2023;25:e45268) doi: 10.2196/45268

KEYWORDS

knowledge graph; natural language processing; neurodevelopmental disorders; autism spectrum disorder; intellectual disability; attention deficit hyperactivity disorder; named entity recognition; topic modeling; aggregation operator

Introduction

Access to curated medical information has become more important than ever due to the growing amount of information available on the internet and the many challenges faced with sharing information about medical topics. Neurodevelopmental disorders (NDDs) are a range of conditions, including autism spectrum disorder, intellectual disability, and attention deficit hyperactivity disorder. These disorders affect up to 18% of the population [1-7] and are influenced by the growing amount of online information and misinformation [8,9]. NDDs have complex medical features, and the needs of affected individuals and their families tend to be quite diverse [10-12].

There exists a large amount of information relating to NDDs on the internet, but this information is scattered across many websites, often using different terminology and containing both reliable information and misinformation. Finding information that is specific, relevant, and trusted is therefore difficult for the caregivers of children with NDDs. To remedy this, a knowledge repository containing available NDD resources annotated with appropriate labels and terms could be constructed. This repository would enable the discovery of relevant trusted resources based on phrases of interest provided by users.

We propose the use of a knowledge graph (KG) to represent web resources together with terms and phrases annotating them. The use of a KG enables web links (ie, resources), terms, and phrases to be represented as nodes, with the relevance between them represented as edges.

A KG indexing web links and information on NDDs would allow experts and nonexperts to have a primary repository of NDD knowledge. With this knowledge, doctors could make quicker and more accurate selections of relevant resources/websites, and caregivers of children with NDDs could quickly find appropriate information, services, and financial support. Accurate identification and early help are critical to quality of life outcomes for those with NDDs. The proposed graph-based repository could improve many peoples' lives.

This paper describes the methodology of constructing a KG-based repository of NDD resources. It presents the following:

- An approach for automatic processing of text extracted from websites relating to NDDs and identification of the most accurate terms/phrases describing them based on named entity recognition (NER), topic modeling, location detection, and resource classification.
- A process of determining degrees of relevance between KG entities and resources, and storing them as weights of relations in the graph.
- An application of the ordered weighted averaging (OWA) operator [13] for determining the most relevant resources using the aggregated weights of relations between resources and terms/phrases describing them.
- An example of using the constructed KG-based repository of NDD resources for retrieving a ranking of resources related to a phrase representing the user's interests.

The paper reviews some related works and describes the methodology used for constructing a KG. It also includes an overview of the KG schema, gives an in-depth look at individual techniques used to annotate scraped web resources, and introduces an aggregation process. Finally, a brief overview of the use of the constructed graph is presented, and an outline of the conclusion and possible future work is provided. KGs have been used in many areas, including medicine, cyber security, finance, news, and education. There have been a wide range of KG applications within the medical field. Applications include general KGs across the whole medical domain and across specific areas, such as depression, thyroid disease, and COVID-19, as described below.

Several KGs spanning the entire medical field have been created. For example, Ernst et al [14] created KnowLife. They used advanced information extraction methods, including NER, pattern mining, and consistency reasoning, to populate entities and relations from scientific literature and online communities, in contrast to many previous works involving manual curation. Shi et al [15] developed methods to extract syntactic, semantic, and structural information from conceptual KGs. They used a similar method to KnowLife for creating the KGs, and extended understanding of the resultant KGs by using machine learning methods to prune meaningless relations in the graphs and extract semantic knowledge. Sheng et al [16] created DEKGB, a KG of various diseases, using prior medical knowledge and electronic medical records, along with guidance from doctors. Li et al [17] used quadruplets instead of triples to represent their KG, with extra information relating to relation strength. Zhang et al [18] used a clinician-in-the-loop approach to fine-tune an automated KG construction method. KGs have also been created for specific medical domains. Huang et al [19] made a KG solely focused on depression after observing the prohibitive size and high-level nature of general medical KGs. A low-level KG for depression would allow more convenient use by doctors, easier understandability by the public, and higher computational efficiency. Chai [20] used a KG about thyroid disease as the backbone for an intelligent medical diagnosis system. Vector embeddings were calculated for each entity and relation in the KG. These embeddings were then used to train a bidirectional long short-term memory (LSTM) network as a disease diagnosis model, outperforming other tested machine learning models. Flocco et al [21] used tweets related to COVID-19 in the Los Angeles area, along with policy announcements and disease spread statistics, to construct a KG representing the real-world spread of COVID-19 in the Los Angeles area. The sentiment of each tweet was calculated using a rule-based method, and topic modeling was used to extract popular keywords from tweets.

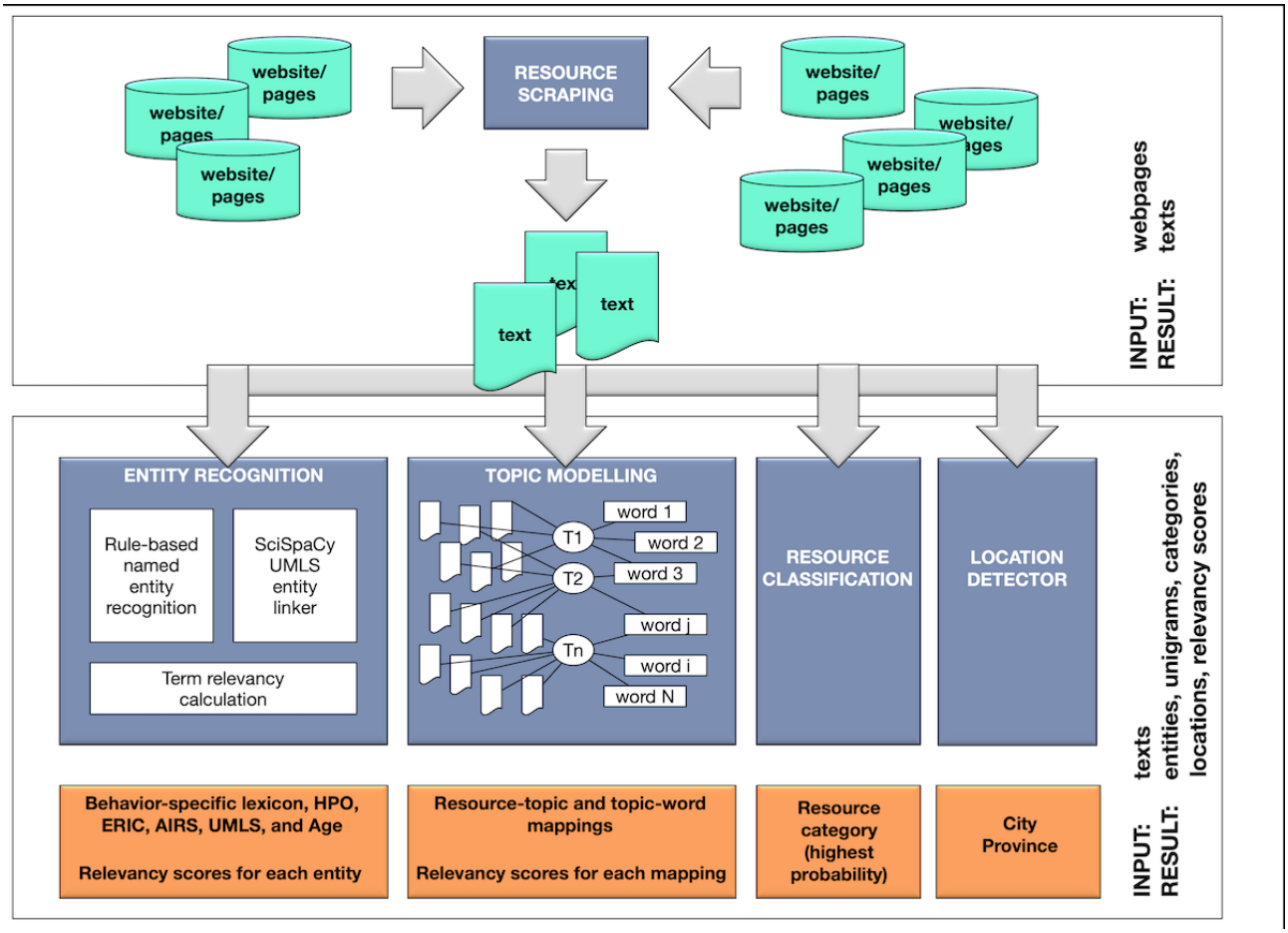
The novelty of our constructed KG lies in domain specificity, the inclusion of patient-focused information from different sources, and the application of combining different information extraction methods. Most other medical KGs focus on the entire medical field and will therefore lose granularity on specific medical topics. We created a KG for NDDs involving input from patients and caretakers affected by NDDs, along with medical professionals who specialize in NDDs. This resulted in a KG containing more extensive knowledge about NDDs than a general medical KG. Input from patients and caretakers allowed us to include resources related to core knowledge, financial help, education, and services. This is in contrast to most other medical KGs, which only focus on extracting medical knowledge from the literature. In addition to the NER pipeline to detect standard terminologies used by medical professionals, we used topic modeling to capture resource-specific keywords. Using both NER and topic modeling allowed us to better annotate the resources. Furthermore, document classification was applied to categorize and label the resources into core knowledge, financial help, education, and services. Representing the extracted knowledge along with the resources in a KG leads to a centralized hub that combines resources from different areas of need around NDDs to maximize knowledge capture.

Methods

Overview

Constructing a KG requires data and methods to represent these data in a format suitable to be a part of the KG. The following sections outline how data were collected and the methodology used to process the data for graph construction purposes. The proposed and applied methodology is illustrated in Figure 1. Data processing methods used to analyze text from websites included NER, topic modeling, document classification, and location detection.

Figure 1. Annotation process of the resource website/pages. AIRS: Alliance of Information and Referral System; ERIC: Education Resources Information Center; HPO: Human Phenotype Ontology; UMLS: Unified Medical Language System.



Data Collection

Two sources were used to construct the NDD corpus of text required for the KG. The first source included individuals with lived experiences who were part of the family advisory board or were recruited through advertisements for the project and community support groups focused on NDDs (AIDE Canada [22] and the Alberta Children's Hospital NDD Care Coordination Project [23]). These individuals/parents were asked to provide links to websites relating to NDDs in such categories as core knowledge, education, services, and funding. The corpus created from these sources is referred to as the *NDD Caregiver subset*.

The second source of relevant web pages used for scraping was the Inform Alberta website [24]. This is referred to as the *Inform Alberta subset*. Finally, the combined list of web pages from both sources and some relevant pages added by the authors were scraped using the Python Scrapy library. For home pages, the

entire site was scraped, while for specific/single web pages, only those pages were scraped.

As a result, the obtained corpus consisted of 200,000 web pages, with 80,000 pages from the *NDD Caregiver subset* and 120,000 pages from the *Inform Alberta subset*. HTML text was extracted for each page and cleaned by removing boilerplate text using the Python BoilerPy3 library. The collection of cleaned HTML text from the web pages formed the corpus of documents used for the construction of the KG-based repository of NDD resources.

NER Approach

The list of NDD resources contained a mixture of website, home page, and web page URLs. To perform web page–level indexing, when a given URL referred to a home page/website, the Scrapy framework was used to scrape all the web pages of that website. Repetitive URLs were removed from the final list of all the web pages. Many web pages contained the same HTML boilerplate, such as headers, navigation bars, and footers. The Boilerpy3 Python library was used to remove this boilerplate HTML.

Entity Vocabulary

The data set contained various web pages related to services, education, financial help, and core health knowledge within the NDD field. Different standard terminologies, including those related to the Unified Medical Language System (UMLS), Human Phenotype Ontology (HPO), Education Resources Information Center (ERIC) thesaurus, and Alliance of Information and Referral Systems (AIRS) taxonomy, were extracted from the pages. In the constructed graph, they were used to annotate the web page URLs.

The UMLS is a collection containing over 4 million concepts from over 100 controlled vocabularies including but not limited to ICD-10 (International Classification of Diseases, Tenth Revision), MeSH (Medical Subject Headings), and SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) [25]. It covers all medical and related entities. The HPO provides a standardized vocabulary of phenotypic abnormalities encountered in human diseases. It currently contains over 13,000 terms [26]. The ERIC thesaurus is a list of topics in education and comprises about 11,818 terms, including 4552 unique terms called descriptors and 7133 synonyms of descriptors [27]. The AIRS taxonomy is the North American standard for indexing and accessing human service resource databases [28]. The taxonomy is a hierarchical system containing more than 9000 terms covering the complete range of human services.

As some web pages were more specific to a particular age as well as location, a list of age terms and all Canadian cities and provinces was used to index web pages. As behavioral issues are common in individuals with NDDs, with expert feedback, the following 10 categories of challenging behaviors were considered: sleep issues, sensory issues, hyperactivity, inattention, repetitive behavior, speech and language development, adaptive behavior, cognitive development, social skills, and behavioral concerns. For each category, we collected commonly used phrases or synonyms with the help of the parent advisory group, as well as performed a manual search on the UMLS interface [29].

NER Process

NER is a subtask of natural language understanding used to detect named entities that refer to specific objects. The named entities we used were domain-specific terms such as medical terms, educational terms, services, challenging behaviors, age, and location. All controlled vocabulary terms were given an entity label the same as their source vocabularies (ie, HPO, ERIC, AIRS, age, and location). Similarly, all challenging behavior phrases or vocabularies were labeled with their respective categories. They were lemmatized using the NLTK library [30]. A single pattern file was created as an input into SpaCy's rule-based entity recognition component called EntityRuler. A pattern file is a dictionary with 2 keys: a "label" specifying the label to be assigned to the entity if the pattern is matched, and a "pattern" indicating the phrase to be matched. Web page text was preprocessed by removing stop words and lemmatizing the text, and was passed to EntityRuler to annotate the text. The UMLS Entity Linker from an open-source framework SciSpaCy [31] was used to extract UMLS entities from the text, and only the respective canonical concepts of UMLS entities were considered for further analysis.

Entity Relevance Calculation

Indexing the web pages with the existence or nonexistence of an entity does not provide information if a document is more relevant to a given entity. A document that mentions a given entity more often than other documents could be considered more relevant to this entity. Depending upon the number of occurrences of an entity, a weight is assigned to each entity, which is called an *entity relevance weight*. The weight provides information on how relevant an entity is to a document. However, using the term (entity) frequency alone will favor common words as well as long documents [32].

It is essential to normalize the term (entity) frequency to incorporate such factors as high term frequency and document length. This is especially so in the case of HTML documents because of keyword stuffing, a process where website owners deliberately add specific keywords to their site in order to improve its search engine ranking. We used logarithmic term frequency as a way to de-emphasize high-frequency terms and adjust within-document term frequency.

For normalization, the *pivoted unique normalization* method was used, which considers the document length as a factor. The principle of the pivoted normalization is as follows: the higher the value of the normalization factor for a document, the lower the chances of its retrieval. Therefore, to boost the chances of retrieving documents of a certain length, the normalization factor for those documents should be lowered. Singal et al [32,33] suggested considering the average document length in a corpus as a reference point, called *pivot*, and using a parameter called *slope* to penalize longer documents and give higher weight to shorter documents. Normalized term relevancy weight is defined as follows:

$$\text{relevance} = 1 + \log(\text{tf}) / (1-\text{slope}) \times \text{pivot} + (\text{slope} \times \text{dl}) \quad \textbf{(1)}$$

where *tf* is the term frequency in the document, and *slope* is set to 0.2 as suggested in the work by Singal et al. The value of *pivot* is set to the average number of distinct named entities per document in the entire collection, and *dl* is the length of the documents referred to by the unique number of entities in the documents. Documents with *dl* = *pivot* are not penalized as the normalization factor is simply equal to the *pivot*. For *dl* > *pivot*, documents are penalized and have lower chances of retrieval, while for *dl* < *pivot*, documents are rewarded with a smaller normalization factor.

Topic Modeling

Topic modeling using latent Dirichlet allocation (LDA) was used to extract similar topics across the corpus for inclusion in the KG. A novel form of topic modeling, referred to as hierarchical topic modeling (HTM), was used to extract more specific topics from the corpus. Topic modeling was performed separately on the *NDD Caregiver subset* and on the *Inform Alberta subset* of the corpus due to computational constraints. Unigram topics were extracted.

Data Preparation

Each web page (document) in a subset of the corpus was preprocessed before being transformed into a count vector for modeling with LDA. The first step was to remove all punctuations from the document, followed by changing all words to lowercase. Next, the document was tokenized and lemmatized. Finally, a stop list was used to remove unwanted words from the document. The stop list used here was the default English stop list from NLTK augmented with some words added by the authors through iterative testing and analysis of the topic modeling outputs. Finally, preprocessed documents were transformed into a count vector for LDA.

Process Description

The HTM algorithm initially performed LDA on the corpus subset and reperformed LDA on topics containing several documents greater than a chosen threshold. Then, the process was repeated until each topic included less than the threshold number of documents, or no more progress was made. It resulted in more specific topic words than running LDA once over the whole corpus, as found by a subjective analysis comparing the outputs of both methods. The LDA algorithm from the Python scikit-learn library was used with the following hyperparameters: maximum iterations of 10, online learning algorithm, learning decay of 0.7, batch size of 128, and maximum features of 50,000.

For the *NDD Caregiver subset* of the corpus, the initial LDA was set to have 200 topics, and for the *Inform Alberta subset*, it was set to have 300 topics. These numbers were chosen to be in proportion to the number of documents in each corpus subset. The threshold for hierarchy termination was set to 300 documents for both corpus subsets. Only the lowest level of the topic hierarchy was used for the KG construction, as these topics seemed to be the most relevant following a subjective analysis.

Topic Relevance Calculation

It is essential to have information about the "strength" of connections among identified topics, documents (web pages), and unigrams, that is, words identified by LDA as describing each topic and indirectly representing documents associated with a given topic. In the case of LDA, such information was extracted from the LDA algorithm.

Document Classification

There were 5 categories of web pages in the corpus: "financial help," "education," "services," "core knowledge/health," and "other." To automatically label each web page, a few classification models were investigated. To construct models, a subset of the corpus was hand labeled as belonging to one or more of the 5 categories. This is a multilabel classification task, as documents (web pages) can belong to more than one category.

The hand-labeled data consisted of 2158 documents, with 116 labeled as "financial help," 420 as "education," 1419 as "services," 1024 as "core knowledge/health," and 143 as "other." This data set was highly unbalanced. The data set was split into training, validation, and testing sets, with 80% of the data used for training, 10% for validation, and 10% for testing. The data were split equally along categories where possible.

We tested 3 groups of models for classifying these documents: (1) multilabel k-nearest neighbors, (2) 5 single-label transformers, and (3) a multilabel transformer. Among these models, we ultimately chose the multilabel transformer, as it achieved the highest macro F1 score on a held-out test set. The multilabel transformer was a 6-layer version of MiniLM-v2 fine-tuned on the prepared training data set. The pretrained model found on the HuggingFace website named nreimers/MiniLM-L6-H384-uncased was used; it is the same model as the *all-MiniLM-L6-v2* from *Sentence-BERT* [34]. A dropout layer with a dropout probability of 0.3 and a final sigmoid activation layer with 5 outputs were added to the base model as a multilabel classification head.

Training hyperparameters for this model were as follows. The loss function used for training was the binary cross entropy loss that was optimized using the AdamW optimizer. The optimizer learning rate was 3×10^{-4}, with a linear warmup to this value and cosine decay to one-tenth of this value during training. The other optimizer hyperparameters were β_1=0.9, β_2=0.95, $=1\times10^{-8}$, and weight decay=0.01. The batch size was 64, and all gradients were clipped to a norm of 1.0 to mitigate gradient explosion.

The model was fine-tuned for 20 epochs. This is higher than the 2 to 3 epochs used in the original BERT paper [35], but we mitigated possible overfitting by increasing the dropout probability and evaluating model performance on a held-out validation set after each epoch. The model with the best performance on the validation set was chosen as the final model. The model outputs 5 probabilities between 0.0 and 1.0. A threshold value was chosen, where values above this threshold were considered members of the corresponding class. Finally, a more fine-tuned macro F1 score was calculated on the validation data set for threshold values from 0.0 to 1.0 in intervals of 0.1.

The best version of the multilabel transformer model, determined based on the validation set, achieved a macro F1 score of 0.504

and an accuracy of 84.1% on the testing set. For reference, the training set macro F1 score was 0.804, with an accuracy of 93.8%. Details of selected model performance can be found in Multimedia Appendix 1.

Location Detection

Using regular expressions, link text was matched to scrape specific pages such as "contact us," "our locations," and "locate us." Then, Canadian/US postal codes were matched using regular expressions and queried using the Google Maps application programming interface to get the city and province for a given postal code. Named entities were detected, along with cities and provinces. To get the final annotations, results from both modules were combined. As it was challenging to remove false-positive location entities due to manual annotation requirements, each city/province was given a weight equal to the proportion of entities that refer to the city/province. This way, for a given city/province, resources could be ranked based upon the score.

Ethical Considerations

This project was approved by the research ethics board at the University of Alberta (study ID: Pro00081113). All the web pages that were analyzed are in the public domain.

Results

Overview

The presented methodology of processing resources (ie, web pages) provides a collection of items (ie, entities, unigrams, age ranges, locations, and web page categories) and challenging behaviors used to annotate the web pages. The integration of this information was done using a KG (KG-based repository of NDD resources). The web pages and items mentioned above were nodes, while the relevance between them was represented as edges labeled with a relevancy strength.

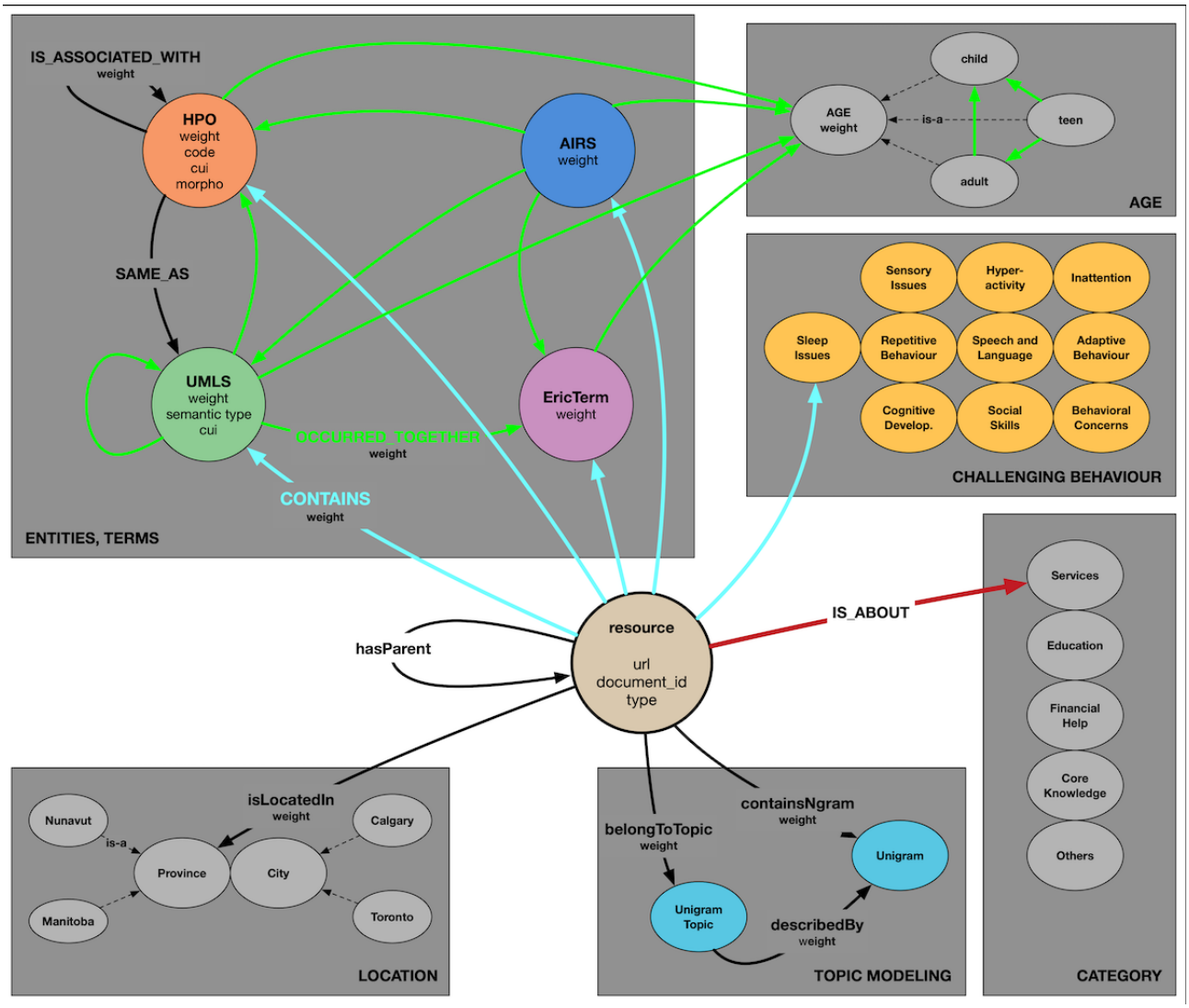
KG Schema

A KG-based repository of NDD resources, as any KG, is a network of entities connected through relations. Each piece of knowledge in a KG is represented as a triple, with 2 entities connected through a relation. These triples are in the form of *(subject, relation, object)*. For example, to represent the piece of information that Edmonton is the capital of Alberta in a graph, the following triple is used: *(Edmonton, capital of, Alberta)*.

To effectively use a KG, names representing types of KG nodes and relations between the nodes must be established. This set, called *vocabulary*, is one of the essential aspects of constructing a KG. The vocabulary is often called the KG schema.

The KG-based repository of NDD resources schema is shown in Figure 2. Names of node types are represented by circles, differing in color by node type. Some of them are labeled with extra information, shown in the text inside the node. Links between the nodes represent relations between entities. Some relations are labeled with extra information, shown in the text on the relation arrow. Each node type and relation type are outlined in the following sections. All collected data were represented as triples and fed into Neo4j to construct the graph automatically.

Figure 2. Knowledge graph schema. Links of the same color represent the same relations; dashed links represent the relation "is-a." AIRS: Alliance of Information and Referral System; ERIC: Education Resources Information Center; HPO: Human Phenotype Ontology; UMLS: Unified Medical Language System.



Entities as Nodes

There were a total of 11 node types in the KG-based repository of NDD resources. The primary node type was *Resource*. It represented all the documents (web page URLs) in the corpus. Each of these resources was labeled with the associated URL from the corpus, the resource source (*NDD Caregiver* or *InformAlberta*), and the resource type (web page, video, or PDF). The document text was not saved in the KG for size reasons. Instead, an external file was kept with processed document text for each URL, and each URL could be accessed through the internet, assuming the web page was still active.

The other node types were linked by edges, either directly or indirectly, to a *resource* node. HPO, UMLS, EricTerm, AIRS, and challenging behavior nodes were extracted using NER methods. *EricTerm* nodes were from the ERIC database and were labeled by the canonical term for the recognized entity. *UMLS* nodes were from the UMLS database and were marked by the canonical term for the recognized entity, its semantic type, and concept unique identifiers. *HPO* nodes were from the HPO-DDD database and were labeled by the canonical term for the recognized entity, and unique human phenotype and concept unique identifiers. *AIRS* nodes were from the AIRS database and were labeled by the canonical term for the recognized entity.

Two types of nodes represented terms extracted from the topic modeling method. First, unique topic nodes were placed into the KG-based repository of NDD resources, and then, each resource (web page) was linked with the unigram topic. Further, each unique topic node was connected with the corresponding unigram terms that were represented as nodes in the KG-based repository of NDD resources.

The remaining node types were *province*, *city*, *age*, *category*, and *challenging behavior*. The *province* and *city* for each resource were extracted using methods outlined in the location detection method. The *age* associated with each resource was also extracted using similar methods. Possible subtypes for *age* are *child*, *teen*, and *adult*. The node type *category* had 5 subtypes: *services*, *education*, *financial help*, *core knowledge*, and *other*. Resources were linked to one of the subtypes after the classification method outlined in document classification was executed.

Relations and Weights

Nodes were connected by edges to one of eight types of relations. Web pages scraped from a parent website and represented as resource nodes were connected to the relation *hasParent*. Location nodes, for both cities and provinces, were connected to the relation *isLocatedIn*. *City* nodes were connected to their corresponding *province* nodes with the relation *inProvince*. NER-related nodes, *age* nodes, and *challenging behavior* nodes were connected to corresponding resource nodes with the relation *CONTAINS*. NER-related nodes were also connected to identically named entities from different databases with the *IS_ASSOCIATED_WITH* relation. NER-related nodes and *age* were connected to each other with the relation *OCCURRED_TOGETHER*.

Topic nodes were connected to corresponding resource nodes with the relation *belongsToTopic*. The relation *describedBy* was used to connect topic nodes to their contained topic word (unigram nodes). Finally, resources were directly connected to relevant unigrams with the *containsNgram* relation.

Relations were assigned a weight using various methods if applicable. The *inProvince*, *hasParent*, and *describedBy* relations had no weights. The relations *CONTAINS*, *IS_ASSOCIATED_WITH*, and *isLocatedIn* were weighted using term relevancy as outlined in the NER method. The relation *OCCURRED_TOGETHER* was labeled with several co-occurrences of connected entities (nodes).

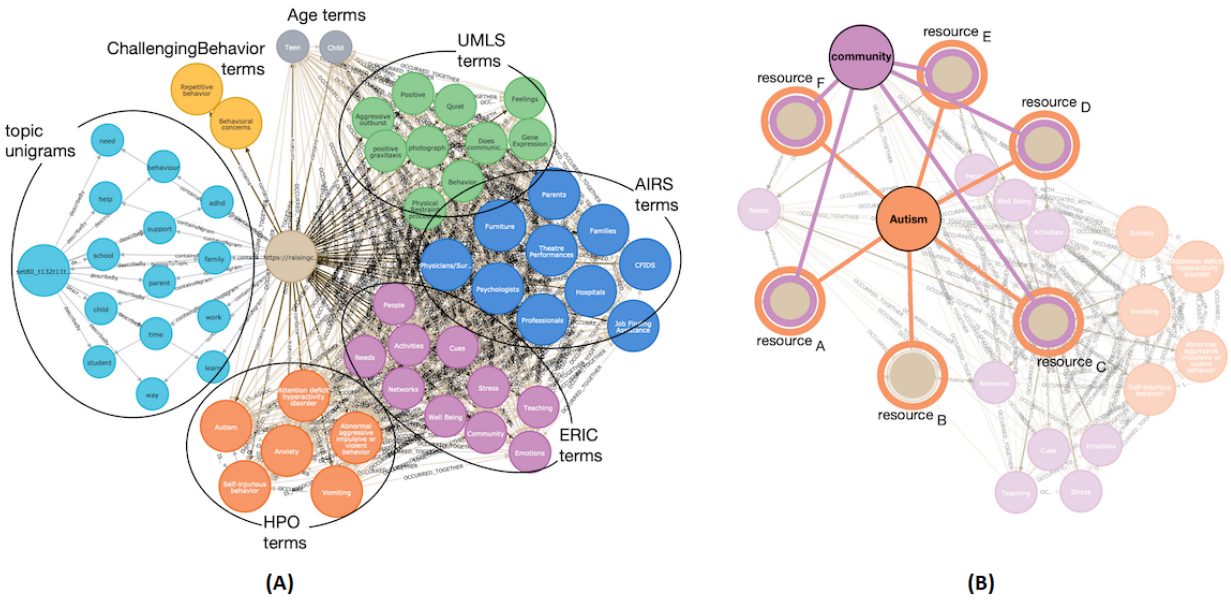
The relations *belongsToTopic*, *containsNgram*, and topic-related *describedBy* had weights calculated as the output of the LDA process. Weights for the *belongsToTopic* relation represented the degree of how strongly a resource belongs to each topic. Weights for the topic-related *describedBy* relation indicated how strongly each word in the topic vocabulary belongs to each topic. The *containsNgram* weights were obtained by matrix multiplication of the *belongsToTopic* and topic-related *describedBy* weight matrices.

Constructed KG: Overview

The constructed KG-based repository of NDD resources contained 264,167 nodes. There were 185,986 resource nodes. For the NER-related approach, there were 2448 *AIRS* nodes, 11,617 *EricTerm* nodes, 4181 *HPO* nodes, and 41,599 *UMLS* nodes. For topic modeling, there were 14,373 unigram nodes and 2045 unigram topic nodes. In addition, there were 3 *age* nodes, 5 *category* nodes, 10 *challenging behavior* nodes, 1832 *city* nodes, and 68 *province*/*state* nodes. The graph contained a total of 22,621,522 relations.

To illustrate interesting features of the graph, a single resource was extracted from the KG-based repository of NDD resources together with several annotated nodes (Figure 3A) [36]. The resource (light brown circle in the middle) is linked with a group of unigrams (blue circles on the left); 2 types of *challenging behavior* nodes (yellow circles); and *age* (gray), *UMLS* (green), *AIRS* (blue), *EricTerm* (violet), and *HPO* (orange) nodes.

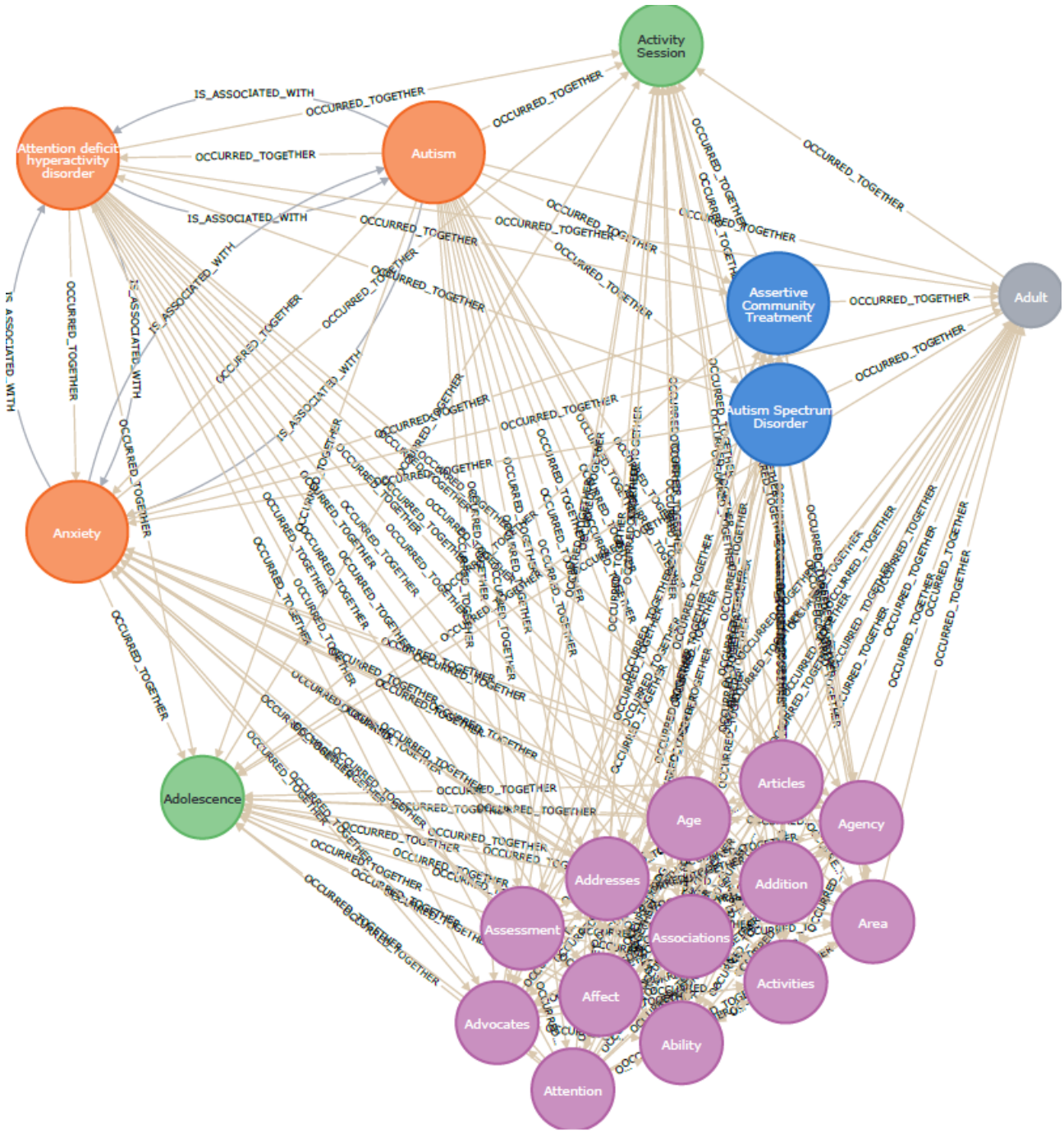
Figure 3. Example of an annotated resource: (A) most relevant annotating nodes; (B) n-to-n relations between resources and annotating nodes. AIRS: Alliance of Information and Referral System; ERIC: Education Resources Information Center; HPO: Human Phenotype Ontology; UMLS: Unified Medical Language System.



These terms and unigrams define/describe the resource. There were n-to-n relations between resources and annotating nodes, meaning that a single annotating node was also linked with multiple resources. Such a scenario has been illustrated in Figure 3B. Two terms (HPO's *autism* and EricTerm's *community*) were connected to multiple resources.

Besides the relations between resources and annotating nodes, the graph contained multiple relations between annotating nodes. These were 2 types of relations (*OCCURRED_TOGETHER* and *IS_ASSOCIATED_WITH*). A fragment of the graph illustrating these relations is shown in Figure 4.

Figure 4. Example of the OCCURRED_TOGETHER and IS_ASSOCIATED_WITH connections between nodes. Entities from different sources are represented in different colors as follows: HPO, orange; AIRS, blue; UMLS, green; ERIC, purple; and age, grey.

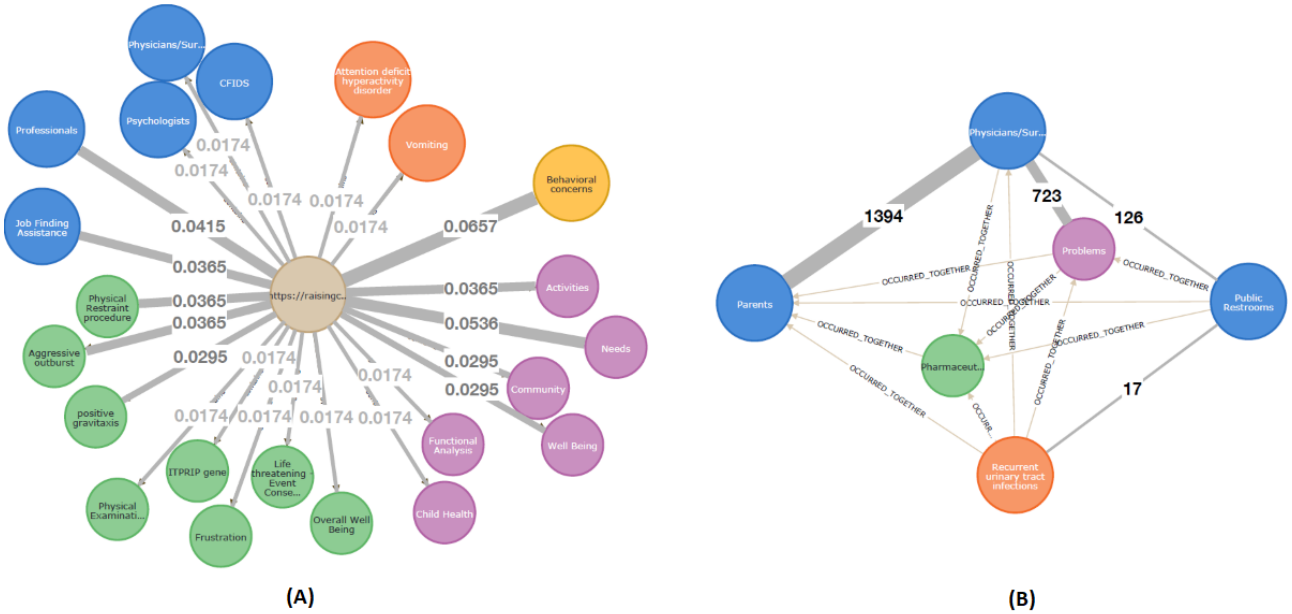


The existence of these many relations can create a highly interconnected representation of resources and their annotated nodes. However, it can introduce an issue if a user tries to identify the most relevant resources, as there would be no difference in relevance between nodes. Therefore, a degree of relevance was added to denote the most appropriate resources. The relevance represented the connection strength between a resource and an annotating node (entity).

Two types of relevance weights were used in the KG-based repository of NDD resources. One weight was linked with the relation *CONTAINS*. Its value was determined using the procedure presented in the entity relevance calculation method. The other weight was linked with the relation *OCCURRED_TOGETHER*. This weight is a measure of the co-occurrence of different annotating nodes.

The first type of weight has been illustrated in Figure 5A. Although all connected nodes contribute to the description of a resource, their contributions are of different strengths. The second type of weight has been illustrated in Figure 5B. A snippet of the graph shows connections between annotating nodes (*HPO*, *AIRS*, and *EricTerm*). The weights were represented as integer numbers that indicated how often both terms co-occurred in the extracted text (ie, degrees to which the given nodes were “related to each other”).

Figure 5. Example of the connection strength of (A) the relation CONTAINS between resources and annotating nodes and (B) the relation OCCURRED_TOGETHER between annotating nodes.



Use of the KG in a User Interface for Resource Retrieval

The KG-based repository of NDD resources can be used to identify the most relevant resources when a user provides some text input. A simple web-based interface has been developed to enable users to use the KG-based repository of NDD resources when they want to obtain a list of relevant resources. The interface allows the user to enter a text query containing several phrases representing their interest. Considering that end users can search with verbose queries, such as “my child hits other children at school,” to infer one of the 10 challenging behavior categories, we have also trained a text classification model that will classify the intent of the entered user text (Multimedia Appendix 2 [35,37-39]). Additionally, the richness of connections of the graph and the need to provide users with the best possible match to their text query led to the need for an aggregation operator that combined the weights (ie, the values of relevance between resources and annotating nodes) in a controlled way. This meant that if a user wanted to find the most relevant resource for a text query that satisfied multiple nodes, an OWA aggregation function was invoked. OWA combines all suitable weights for these nodes to rank the resources based upon their relevance (Multimedia Appendix 3 [13,40]).

The entered text is processed with the following steps:

1. Extract unigrams and entities (HPO, UMLS, ERIC, AIRS, and challenging behavior) using the developed natural language processing pipeline.
2. Classify user text as one of the 10 challenging behavior categories using the transfer learning–based text classification model. Then, add the detected category to the entities list obtained in step 1.
3. Query the KG-based repository of NDD resources to retrieve all resources that are connected to nodes representing entities and unigrams obtained in the above steps 1 and 2. For each retrieved resource, all annotating nodes are extracted together with the weights of the relations.
4. The weights are aggregated using OWA to determine the relevance of each retrieved resource.
5. A list of “sorted by relevance” resources is displayed to the user.

The text query, along with extracted entities and unigrams, is shown for a simple example in Figure 6. As can be seen, *HPO* and *UMLS* entities have been identified (“abnormal aggressive impulsive or violent behavior” and “spitting,” respectively). Additionally, *behavioral concerns* as a category of *challenging behavior* has been recognized. Two unigrams (“aggressive” and “behavior”) are extracted from the text. The resulting list of the most relevant sites has been determined and is shown in Figure 7. The list contains web pages and a video, all from the category *core knowledge*.

Figure 6. Question interface for query matching. The user enters the text "aggressive behavior and kicking and spitting" and obtains entities and unigrams.

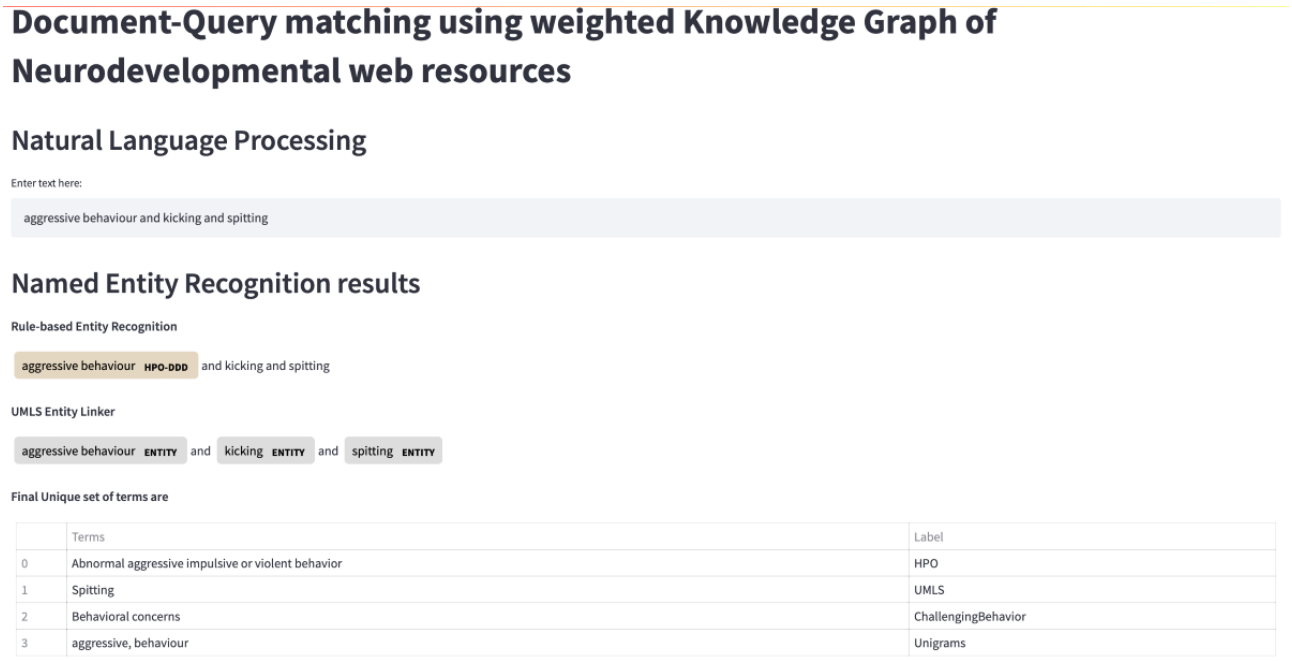
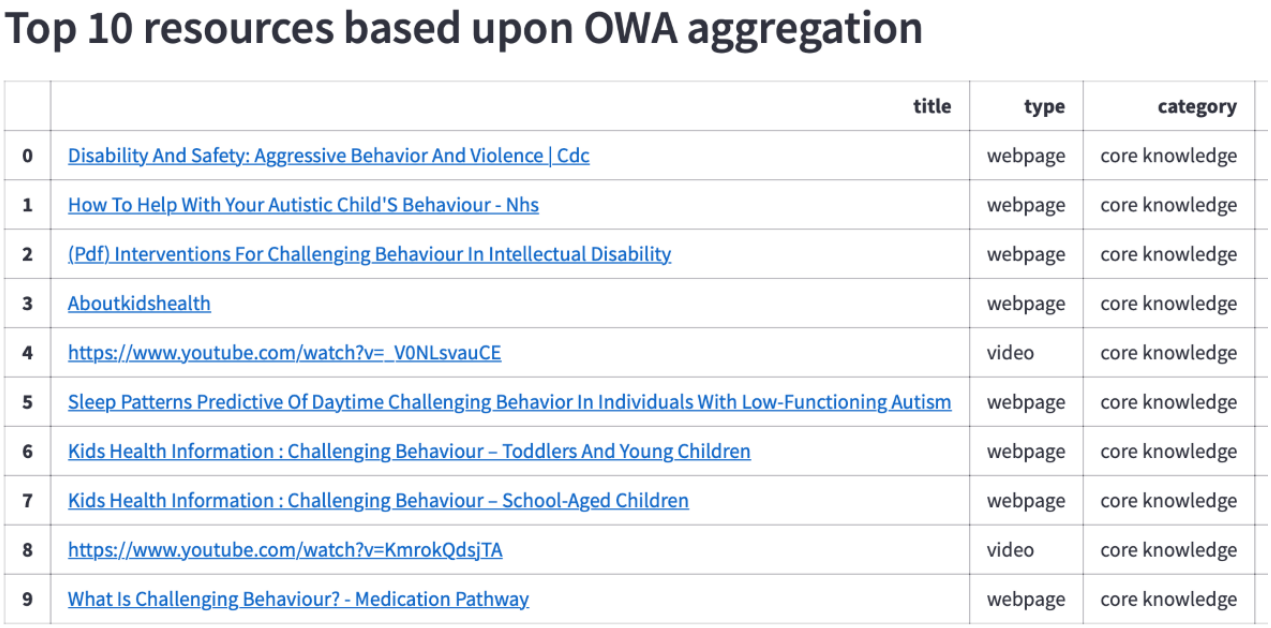


Figure 7. Question interface for the top 10 resources. List of the top 10 most relevant resources for the text "aggressive behavior and kicking and spitting." OWA: ordered weighted averaging.



	title	type	category
0	Disability And Safety: Aggressive Behavior And Violence | Cdc	webpage	core knowledge
1	How To Help With Your Autistic Child'S Behaviour - Nhs	webpage	core knowledge
2	(Pdf) Interventions For Challenging Behaviour In Intellectual Disability	webpage	core knowledge
3	Aboutkidshealth	webpage	core knowledge
4	https://www.youtube.com/watch?v=_V0NLsvauCE	video	core knowledge
5	Sleep Patterns Predictive Of Daytime Challenging Behavior In Individuals With Low-Functioning Autism	webpage	core knowledge
6	Kids Health Information : Challenging Behaviour – Toddlers And Young Children	webpage	core knowledge
7	Kids Health Information : Challenging Behaviour – School-Aged Children	webpage	core knowledge
8	https://www.youtube.com/watch?v=KmrokQdsjTA	video	core knowledge
9	What Is Challenging Behaviour? - Medication Pathway	webpage	core knowledge

Discussion

This paper describes the methodology for processing text extracted from web pages to generate a set of entities and terms used for annotating these web pages. Both web pages and entities/terms are the basis for constructing a knowledge base of resources. This base, built as a graph called the KG-based repository of NDD resources, is a highly interconnected network linking resources with annotating terms and entities. Furthermore, edges in the KG-based repository of NDD resources have weights representing relevance between resources and annotating terms/entities. Edge weights, aggregated using a specialized aggregation operator, are used to rank resources. The constructed KG-based repository of NDD resources is a repository of resources about NDDs that can be queried using textual phrases, with relevant results shown to the user using an interface.

Most of the prior work in building a medical KG has used scientific literature, such as PubMed and electronic medical records, and only specific types of entities, such as diseases, chemicals, and genes, have been considered [17,41,42]. Ernst et al [43] used patient-oriented online health portals to build a KG, indicating the importance of medical information spread across different sources. Shi et al [15] represented heterogeneous

textual medical knowledge as a KG to use it further for semantic reasoning. Yu et al [44] constructed a KG for traditional Chinese medicine to integrate terms, documents, and databases in a base to facilitate sharing and use of traditional Chinese medicine health care knowledge.

To our knowledge, this is the first method to integrate credible online information from different areas of need around NDDs (ie, financial help, services, education, and core knowledge) into one base. Our developed natural language processing pipeline can be used to annotate resources from the abovementioned areas. Representing the extracted knowledge in a KG allows for finding connections among different resources on a scale that would be impossible for a single human. Many ontologies and scattered information at both professional and layperson levels exist on the internet, and our KG compiles all this information in a single place. Connecting all this information will open up many areas of improvement for the NDD field. These could include new research directions, new treatment opportunities, and the possibility of collaboration among services.

The methodology used to construct this KG is scalable and could be expanded to other medical domains besides NDDs. In creating more of these domain-specific medical KGs with the guidance of medical professionals, patients, and caretakers, we can provide information in a similar way to patients having other conditions. These specific KGs could even be connected at a higher level to slowly create a field-wide medical KG, which would be of great benefit in complex medical conditions where individuals present with multiple hyperspecialized domains. The bottom-up nature of the creation of this KG may result in a better product than the top-down field-wide medical KGs that currently exist.

While having more documents means more information is available, there is also a tradeoff between the number of documents and the KG query speed [45]. Some ways to overcome this include indexing the KG [46], and pruning irrelevant nodes and edges on the KG [47]. Another limitation of KG construction is that pivoted unique normalized logarithmic term frequency is used to calculate weights for edges labeled *CONTAINS*, which can affect the performance of the resource retrieval method when the size of the document is significantly greater than the average document length in the corpus. Pivoted unique normalization overpenalizes longer documents as shown in the original paper [48]. When the length of a document is much larger than the average document length in the corpus, a higher normalization factor could yield an almost zero relevance score for that document's entities [48]. This limitation can be overcome by implementing a term relevance method, which considers not only the term frequency but also co-occurring terms (represented with the *OCCURRED_TOGETHER* relationship in the KG-based repository of NDD resources) [49]. As future work, the semantic search methods of the KG-based repository of NDD resources will be further studied to address the exact keyword matching issue in the resource retrieval system. Potential solutions include using query expansion techniques [50-52]. The application of OWA for identifying the most relevant list of resources opens another possibility of enhancing the user query interface. OWA is known for its ability to include linguistic quantifiers, such as SOME, MOST, ALL, and AT LEAST n, in the aggregation process. So far, we have only used MOST to aggregate query results, yet a user can control to what degree documents should satisfy different criteria using different quantifiers.

To further improve the user's experience with the resource retrieval process of the KG-based repository of NDD resources, we aim to build a transparent interface that will enable path-based explanations in the KG-based repository of NDD resources to provide relevant background knowledge in a human-understandable format [53,54], using interpretable machine learning approaches. Explainable artificial intelligence is an emerging research field, which focuses on not only the performance of the models but also the interpretability of what factors led the model to make a particular decision. This promotes credibility and trust in the suggested results [55,56].

Although patients and caretakers were included in some of the vital steps of creating this KG, such as collecting resources and challenging behavior vocabulary, user feedback is an important step in the process of validating our created KG. We will design an evaluation strategy to validate the document retrieval system of the KG-based repository of NDD resources by collecting a gold standard relevance assessment from human judges. Douze et al [57] found that the relevance assessment from human judges depends upon their subjective needs. Therefore, we will collaborate with a group of parents of children with NDDs to create a gold standard test collection to evaluate the model and check if the document retrieval system of the KG-based repository of NDD resources satisfies their needs.

The need for helping families with NDDs could leverage the potential that online information has to offer (eg, to supplement gaps in the health/social support system). This need became more important than ever during the COVID-19 pandemic and will continue to gain importance in the future. Building an efficient repository of trusted web resources has proven to be challenging due to the lack of uniformly labeled resources. This challenge is not unique to NDDs and is seen across other medical fields as well. Such repositories of online resources should provide users with an intelligently generated ranking of resources based on a simple text query entered by the users. Experts and nonexperts can use the KG-based repository of NDD resources to improve the quality of life of people with NDDs. Future work includes enhancement of the user interface for resource retrieval, as well as mechanisms for continuous modification of the KG-based repository of NDD resources when new information is discovered or old information is found to be outdated.

Acknowledgments

This work was funded by the Canadian Institute of Health Research (CIHR) and the Natural Science and Engineering Research Council of Canada (NSERC). We would like to thank Cory Rosenfelt for assistance in manuscript editing and Navid Rezaei for implementing the text classification model for the user interface.

Authors' Contributions

FVB and MZR conceptualized the project. MZR designed the methodology. JC and MK implemented the natural language processing pipeline. MK developed the knowledge graph and implemented the resource ranking application. MK and JC implemented the user interface. JC, MK, MZR, and FVB wrote the manuscript. FVB and MZR supervised the project.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Multilabel transformer model performance results.
[DOCX File , 14 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Methodology of the language model–based challenging behavior detection.
[DOCX File , 74 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Ordered weighted averaging–based resource relevance ranking algorithm.
[DOCX File , 17 KB-Multimedia Appendix 3]

References

1. Arora NK, Nair MKC, Gulati S, Deshmukh V, Mohapatra A, Mishra D, et al. Neurodevelopmental disorders in children aged 2-9 years: Population-based burden estimates across five regions in India. PLoS Med 2018 Jul;15(7):e1002615 [FREE Full text] [doi: 10.1371/journal.pmed.1002615] [Medline: 30040859]
2. Emerson E. Deprivation, ethnicity and the prevalence of intellectual and developmental disabilities. J Epidemiol Community Health 2012 Mar;66(3):218-224. [doi: 10.1136/jech.2010.111773] [Medline: 20889590]
3. Taylor E. Developing ADHD. J Child Psychol Psychiatry 2009 Jan;50(1-2):126-132. [doi: 10.1111/j.1469-7610.2008.01999.x] [Medline: 19076263]
4. Johnson S, Fawke J, Hennessy E, Rowell V, Thomas S, Wolke D, et al. Neurodevelopmental disability through 11 years of age in children born before 26 weeks of gestation. Pediatrics 2009 Aug 27;124(2):e249-e257. [doi: 10.1542/peds.2008-3743] [Medline: 19651566]
5. Zauche LH, Darcy Mahoney AE, Higgins MK. Predictors of Co-occurring Neurodevelopmental Disabilities in Children With Autism Spectrum Disorders. J Pediatr Nurs 2017;35:113-119. [doi: 10.1016/j.pedn.2017.04.002] [Medline: 28728761]
6. Hansen BH, Oerbeck B, Skirbekk B, Petrovski BÉ, Kristensen H. Neurodevelopmental disorders: prevalence and comorbidity in children referred to mental health services. Nord J Psychiatry 2018 May;72(4):285-291. [doi: 10.1080/08039488.2018.1444087] [Medline: 29488416]
7. Tatishvili N, Gabunia M, Laliani N, Tatishvili S. Epidemiology of neurodevelopmental disorders in 2 years old Georgian children. Pilot study - population based prospective study in a randomly chosen sample. Eur J Paediatr Neurol 2010 May;14(3):247-252. [doi: 10.1016/j.ejpn.2009.07.004] [Medline: 19683948]
8. Zucker HA. Tackling Online Misinformation: A Critical Component of Effective Public Health Response in the 21st Century. Am J Public Health 2020 Oct;110(S3):S269. [doi: 10.2105/AJPH.2020.305942] [Medline: 33001725]
9. Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. Information Processing & Management 2021 Jan;58(1):102390. [doi: 10.1016/j.ipm.2020.102390]
10. Parker M, Killian M. Autism spectrum disorder and complex healthcare needs: The role of healthcare experiences. Research in Autism Spectrum Disorders 2020 May;73:101535. [doi: 10.1016/j.rasd.2020.101535]
11. Eklund H, Findon J, Cadman T, Hayward H, Murphy D, Asherson P, et al. Needs of Adolescents and Young Adults with Neurodevelopmental Disorders: Comparisons of Young People and Parent Perspectives. J Autism Dev Disord 2018 Jan;48(1):83-91 [FREE Full text] [doi: 10.1007/s10803-017-3295-x] [Medline: 28894999]
12. Bloch JS, Weinstein JD. Families of Young Children With Autism. Social Work in Mental Health 2009 Dec 11;8(1):23-40. [doi: 10.1080/15332980902932342]

13. Yager RR. On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking. Readings in Fuzzy Sets for Intelligent Systems 1993:80-87. [doi: 10.1016/B978-1-4832-1450-4.50011-0]
14. Ernst P, Meng C, Siu A, Weikum G. KnowLife: A knowledge graph for health and life sciences. 2014 Presented at: 30th International Conference on Data Engineering; March 31, 2014-April 04, 2014; Chicago, IL, USA. [doi: 10.1109/ICDE.2014.6816754]
15. Shi L, Li S, Yang X, Qi J, Pan G, Zhou B. Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services. Biomed Res Int 2017;2017:2858423 [FREE Full text] [doi: 10.1155/2017/2858423] [Medline: 28299322]
16. Sheng M, Shao Y, Zhang Y, Li C, Xing C, Zhang H, et al. DEKGB: An Extensible Framework for Health Knowledge Graph. In: Chen H, Zeng D, Yan X, Xing C, editors. Smart Health. ICSH 2019. Lecture Notes in Computer Science, vol 11924. Cham: Springer; 2019:27-38.
17. Li L, Wang P, Yan J, Wang Y, Li S, Jiang J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med 2020 Mar;103:101817. [doi: 10.1016/j.artmed.2020.101817] [Medline: 32143785]
18. Zhang Y, Sheng M, Zhou R, Wang Y, Han G, Zhang H, et al. HKGB: An Inclusive, Extensible, Intelligent, Semi-auto-constructed Knowledge Graph Framework for Healthcare with Clinicians' Expertise Incorporated. Information Processing & Management 2020 Nov;57(6):102324 [FREE Full text] [doi: 10.1016/j.ipm.2020.102324]
19. Huang Z, Yang J, van Harmelen F, Hu Q. Constructing Knowledge Graphs of Depression. In: Health Information Science. HIS 2017. Lecture Notes in Computer Science, vol 10594. Cham: Springer; 2017:149-161.
20. Chai X. Diagnosis Method of Thyroid Disease Combining Knowledge Graph and Deep Learning. IEEE Access 2020;8:149787-149795 [FREE Full text] [doi: 10.1109/access.2020.3016676]
21. Flocco D, Palmer-Toy B, Wang R, Zhu H, Sonthalia R, Lin J, et al. An Analysis of COVID-19 Knowledge Graph Construction and Applications. 2021 Presented at: 2021 IEEE International Conference on Big Data (Big Data); December 15-18, 2021; Orlando, FL, USA. [doi: 10.1109/BigData52589.2021.9671479]
22. AIDE Canada. URL: https://aidecanada.ca [accessed 2022-12-19]
23. NDD Care Coordination Project. Alberta Health Services. URL: http://fcrc.albertahealthservices.ca/coordination/about/ [accessed 2022-12-19]
24. InformAlberta. URL: https://informalberta.ca/ [accessed 2022-12-19]
25. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
26. Köhler S, Gargano M, Matentzoglu N, Carmody L, Lewis-Smith D, Vasilevsky N, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res 2021 Jan 08;49(D1):D1207-D1217 [FREE Full text] [doi: 10.1093/nar/gkaa1043] [Medline: 33264411]
27. ERIC - Education Resources Information Center. URL: https://eric.ed.gov/?ti=all [accessed 2022-12-19]
28. The Taxonomy. AIRS. URL: https://www.airs.org/i4a/pages/index.cfm?pageid=3386 [accessed 2022-12-09]
29. UMLS Metathesaurus Browser. URL: https://uts.nlm.nih.gov/uts/umls/home [accessed 2023-02-20]
30. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Sebastopol, CA, USA: O'Reilly; 2009.
31. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August 2019; Florence, Italy. [doi: 10.18653/v1/w19-5034]
32. Polettini N. The Vector Space Model in Information Retrieval-Term Weighting Problem. ResearchGate. URL: https://www.researchgate.net/publication/229053068_The_Vector_Space_Model_in_Information_Retrieval-Term_Weighting_Problem [accessed 2023-03-25]
33. Singhal A, Buckley C, Mitra M. Pivoted Document Length Normalization. SIGIR Forum 2017 Aug 02;51(2):176-184 [FREE Full text] [doi: 10.1145/3130348.3130365]
34. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1410]
35. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2019; Minneapolis, Minnesota.
36. Aggressive behaviour: autistic children and teenagers. Raising Children Network. URL: https://raisingchildren.net.au/autism/behaviour/common-concerns/aggressive-behaviour-asd [accessed 2023-04-06]
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: 31st International

Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA, USA. [doi: 10.5555/3295222.3295349]
38. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014;15:1929-1958 [FREE Full text]
39. Decoupled Weight Decay Regularization. Open Review. URL: https://openreview.net/pdf?id=Bkg6RiCqY7 [accessed 2023-04-06]
40. He W, Dutta B, Rodríguez R, Alzahrani A, Martínez L. Induced OWA Operator for Group Decision Making Dealing with Extended Comparative Linguistic Expressions with Symbolic Translation. Mathematics 2020 Dec 23;9(1):20. [doi: 10.3390/math9010020]
41. Gong F, Wang M, Wang H, Wang S, Liu M. SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. Big Data Research 2021 Feb;23:100174. [doi: 10.1016/j.bdr.2020.100174]
42. Wu X, Duan J, Pan Y, Li M. Medical Knowledge Graph: Data Sources, Construction, Reasoning, and Applications. Big Data Min. Anal 2023 Jun;6(2):201-217. [doi: 10.26599/bdma.2022.9020021]
43. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 2015 May 14;16:157 [FREE Full text] [doi: 10.1186/s12859-015-0549-5] [Medline: 25971816]
44. Yu T, Li J, Yu Q, Tian Y, Shun X, Xu L, et al. Knowledge graph for TCM health preservation: Design, construction, and applications. Artif Intell Med 2017 Mar;77:48-52. [doi: 10.1016/j.artmed.2017.04.001] [Medline: 28545611]
45. Big-O Cheat Sheet. static.packt-cdn. URL: https://static.packt-cdn.com/downloads/4874OS_Appendix_Big_O_Cheat_Sheet.pdf [accessed 2023-02-20]
46. Indexes for search performance. Neo4j Graph Data Platform. URL: https://neo4j.com/docs/cypher-manual/5/indexes-for-search-performance/ [accessed 2023-02-20]
47. Jin J, Luo J, Khemmarat S, Gao L. Querying Web-Scale Knowledge Graphs Through Effective Pruning of Search Space. IEEE Trans. Parallel Distrib. Syst 2017 Aug 1;28(8):2342-2356. [doi: 10.1109/tpds.2017.2665478]
48. Lv Y, Zhai CX. Lower-bounding term frequency normalization. In: CIKM '11: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011 Presented at: 20th ACM International Conference on Information and Knowledge Management; October 24-28, 2011; Glasgow, Scotland, UK. [doi: 10.1145/2063576.2063584]
49. Rousseau F, Vazirgiannis M. Graph-of-word and TW-IDF: new approach to ad hoc IR. In: CIKM '13: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013 Presented at: 22nd ACM International Conference on Information & Knowledge Management; October 27, 2013-November 1, 2013; San Francisco, California, USA. [doi: 10.1145/2505515.2505671]
50. Lissandrini M, Mottin D, Palpanas T, Velegrakis Y. Graph-Query Suggestions for Knowledge Graph Exploration. In: WWW '20: Proceedings of The Web Conference 2020. 2020 Presented at: The Web Conference 2020; April 20-24, 2020; Taipei, Taiwan. [doi: 10.1145/3366423.3380005]
51. Xiong C, Callan J. Query Expansion with Freebase. In: ICTIR '15: Proceedings of the 2015 International Conference on The Theory of Information Retrieval. 2015 Presented at: 2015 International Conference on The Theory of Information Retrieval; September 27-30, 2015; Northampton, Massachusetts, USA. [doi: 10.1145/2808194.2809446]
52. Balaneshinkordan S, Kotov A. An Empirical Comparison of Term Association and Knowledge Graphs for Query Expansion. In: Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Cham: Springer; 2016:761-767.
53. Fussl A, Nissen V. Interpretability of Knowledge Graph-based Explainable Process Analysis. 2022 Presented at: Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE); September 19-21, 2022; Laguna Hills, CA, USA. [doi: 10.1109/aike55402.2022.00008]
54. Xian Y, Fu Z, Muthukrishnan S, de Melo G, Zhang Y. Reinforcement Knowledge Graph Reasoning for Explainable Recommendation. In: SIGIR'19: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019 Presented at: 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval; July 21-25, 2019; Paris, France. [doi: 10.17925/ohr.2020.16.1.43]
55. Ammar N, Shaban-Nejad A. Explainable Artificial Intelligence Recommendation System by Leveraging the Semantics of Adverse Childhood Experiences: Proof-of-Concept Prototype Development. JMIR Med Inform 2020 Nov 04;8(11):e18752 [FREE Full text] [doi: 10.2196/18752] [Medline: 33146623]
56. Cheng K, Wang N, Li M. Interpretability of Deep Learning: A Survey. In: Meng H, Lei T, Li M, Li K, Xiong N, Wang L, editors. Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery. ICNC-FSKD 2020. Lecture Notes on Data Engineering and Communications Technologies, vol 88. Cham: Springer; 2021:475-486.
57. Douze L, Pelayo S, Messaadi N, Grosjean J, Kerdelhué G, Marcilly R. Designing Formulae for Ranking Search Results: Mixed Methods Evaluation Study. JMIR Hum Factors 2022 Mar 25;9(1):e30258 [FREE Full text] [doi: 10.2196/30258] [Medline: 35333180]

Abbreviations

AIRS: Alliance of Information and Referral Systems
ERIC: Education Resources Information Center
HPO: Human Phenotype Ontology
HTM: hierarchical topic modeling
KG: knowledge graph
LDA: latent Dirichlet allocation
NDD: neurodevelopmental disorder
NER: named entity recognition
OWA: ordered weighted averaging
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 22.12.22; peer-reviewed by S Chaudhry, Y Wang, M Burns, D Chrimes; comments to author 31.01.23; revised version received 21.02.23; accepted 12.03.23; published 17.04.23

Please cite as:
Costello J, Kaur M, Reformat MZ, Bolduc FV
Leveraging Knowledge Graphs and Natural Language Processing for Automated Web Resource Labeling and Knowledge Mobilization in Neurodevelopmental Disorders: Development and Usability Study
J Med Internet Res 2023;25:e45268
URL: https://www.jmir.org/2023/1/e45268
doi: 10.2196/45268
PMID: 37067865