ArcheGEO: Towards Improving Relevance of Gene Expression Omnibus Search Results

Huey-Eng Chua
Nanyang Technological University
Singapore
hechua@ntu.edu.sg

Lisa Tucker-Kellogg
Duke-NUS Medical School
Singapore
lisa.tucker-kellogg@duke-nus.edu.sg

Sourav S Bhowmick
Nanyang Technological University
Singapore
assourav@ntu.edu.sg

ABSTRACT

Transciptomic data stored in the *Gene Expression Omnibus* (*GEO*) serves thousands of queries per day, but a lack of standardized machine-readable metadata causes many searches to return irrelevant hits, which impede convenient access to useful data in the *GEO* repository. Here, we describe ArcheGEO, a novel end-to-end framework that improves results from the *GEO Browser* by *automatically* determining the *relevance* of these results. Unlike existing tools, ArcheGEO reports on the *irrelevant* results and provides reasoning for their exclusion. Such reasoning can be leveraged to improve annotations of metadata.

ACM Reference Format:
Huey-Eng Chua, Lisa Tucker-Kellogg, and Sourav S Bhowmick. 2022. ArcheGEO: Towards Improving Relevance of Gene Expression Omnibus Search Results. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3535508.3545531

1 INTRODUCTION

The transcriptome is the set of RNA transcripts present in a cell or set of cells. Measuring RNA abundance in different tissue-types, timepoints, or contexts is a popular method to pursue biological questions [31] and is sometimes used for medical questions in human samples [72]. Multiple public repositories of transcriptomic data have been established: *ArrayExpress (AE)* [1], the *Genomic Expression Archive (GEA)* [5], and *Gene Expression Omnibus (GEO)* [4]. In particular, *GEO* is growing rapidly. The growth of *GEO* since its 2020 size [15] has been 64%, with 159481 transcriptome series entry records. Hence, we interpret *GEO* to be the dominant repository.

GEO serves diverse needs throughout the biomedical sciences, and its usage is 15,000 accessions per day by 1000 unique users. Purposes of *GEO* use have been recorded by NCBI (https://www.ncbi.nlm.nih.gov/geo/info/citations.html) with the following most common: (1) discovering function of uncharacterized genes and genetic networks by analyzing of spatial and temporal transcriptomic patterns [34], (2) validating interesting gene expression trends by cross-comparison [33, 66], (3) substantiating experimental discussion and findings [16, 55] and (4) providing insights to related fields (*e.g.,* gene set analysis, cell-type composition analysis) through re-analysis and re-interpretation of *GEO* data [41, 63]. *GEO* currently provides two interfaces for data retrieval, namely, the *GEO Browser* (https://www.ncbi.nlm.nih.gov/geo/browse/) and a suite of programmatic access utilities (https://www.ncbi.nlm.nih.gov/geo/info/geo_paccess.html).

Retrieval of *GEO* entries typically relies on keyword-based search on *GEO* metadata since researchers may not know the exact identifiers of the entries of interest. The lack of standardized machine-readable metadata (*e.g.,* usage of ontologies or thesaurus to describe biological entities) hinders the reuse of *GEO* entries [19, 68, 69] as they can affect the *relevance* of the search results. For example, searching "human breast cancer" in the *GEO DataSets* (https://www.ncbi.nlm.nih.gov/gds) yields 172 results of which 17 are non-human, 20 are non-breast, 18 are non-malignant (Figure 1, last accessed: June 26, 2022). When using *GEO* records for a biologically-defined purpose, users must open each record and inspect the header data (*i.e.,* metadata) to verify compliance with the defined purpose, before proceeding with the use.

Internally, a keyword query on *GEO* is transformed to a search query containing a complex set of attribute-value pairs with AND/OR connectives (*e.g., Search details* text box in Figure 1). It may seem that we can address the aforementioned problem by directly specifying relevant attribute-value pairs in the *advanced search* feature of *GEO*. However, this may not necessarily be the case. For example, searching "(breast cancer[Description]) AND human[Organism]" still retrieves the result related to prostate cancer (Figure 1). Furthermore, the number of search results now decreases to 147. If we use (breast cancer[Title]) AND human[Organism] as a query, the result size further decreases to 113. In addition, it is cognitively challenging for an end user to find the right set of attribute-value pairs and their AND/OR connectives to retrieve all relevant results.

Various metadata-focused approaches have sought to improve *GEO* data reusability and can be broadly classified as (1) manual curation-based [32, 69], (2) automated natural language processing (NLP)-based [19, 27], and (3) gene expression data inference-based [30, 38]. Amongst the three approaches, manual curation produces the best result quality, but requires expert knowledge and is extremely labor-intensive. Although leveraging gene expression data can be used to infer metadata elements such as cell type, organism and platform [68], the amount of information that can be inferred is still very limited and may not yield a rich enough metadata to annotate the *GEO* entries. Furthermore, finding relevant entries by seeking similar expression patterns could lead to confirmation bias in the results. That is, if one searches for entries using gene expression similarity, then they will have improved power to find hits that resemble previous hits (*e.g.,* similar to pre-existing knowledge of gene expression levels for that disease), but may not

BCB '22, August 7–10, 2022, Northbrook, IL, USA

ACM ISBN 978-1-4503-9386-7/22/08…$15.00
https://doi.org/10.1145/3535508.3545531

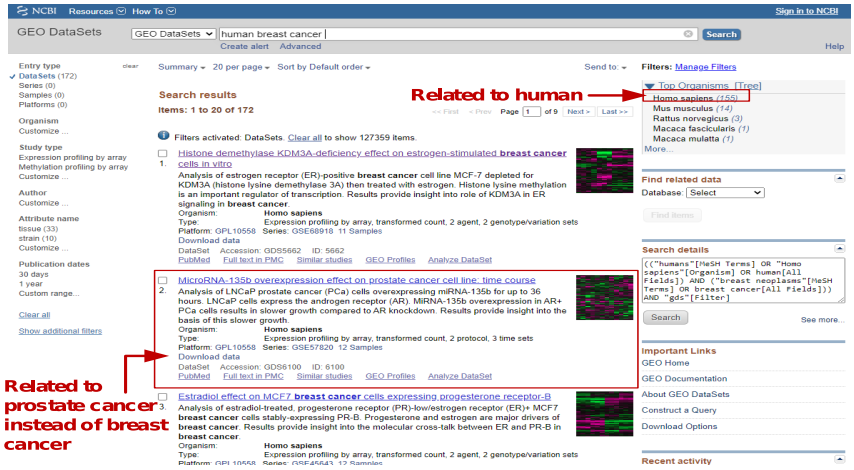

Figure 1: Search results on *GEO*.

have equal improvement at obtaining hits that are dissimilar or different. In comparison, NLP-based approaches can potentially yield a much richer metadata because they extract structured elements from free-text description of the entry provided by the submitter.

Several NLP-based approaches provide annotations that are general descriptions of the entries (*e.g.,* disease and drug annotation [69]). Others provide annotations that facilitate subsequent tasks. Examples include automatic tagging of samples as perturbation or control for comparison analysis [27] and automatic tagging of time series to facilitate dynamic analysis [19]. Such NLP-based approaches complement existing *GEO* interfaces to improve the search result quality for specific research goals [73].

These approaches focus on annotating existing *GEO* records. Search engines such as *GEO Browser* leverage these annotations to generate search results based on user-specified keywords (*i.e.,* queries) and typically work on the assumption that the annotations are *correct*, which may not always be the case. For instance, GDS6100 which is related to prostate cancer instead of breast cancer is erroneously reported as a search result in the *GEO Browser* for query keywords "human" and "breast cancer" (Figure 1). This is because manual curation-based annotation can still be error prone either due to error during data entry of the annotation or insufficient domain knowledge of the curators. Similarly, NLP-based and gene expression data inference-based approaches may yield wrong annotation due to poor quality of training data or poor design of inference algorithm. *Hence, it is paramount to have a verification component to check the relevance of search results in order to improve the search quality.* In this paper, we propose a novel end-to-end framework called ArcheGEO (**A**utomated **R**elevance **CHE**cker for **GEO**), built on top of *GEO*, to realize this component.

Given keywords specified by a user, ArcheGEO *automatically* identifies a set of *irrelevant* matches (if any) in the results returned by the *GEO Browser* and provides reasons for their irrelevance by addressing the *irrelevant match finding* (IMF) problem. For instance, ArcheGEO will identify the second result in Figure 1 as irrelevant and provide reasons for it. Under the hood, it first reconstructs the original query as a set of queries that are disease- and organism-centric. To this end, it first extracts disease-related or phenotype-related *concepts* from user keywords and *GEO* metadata. Then, it leverages controlled vocabularies to determine the *equivalence* of the *concepts*. The reconstructed queries are sent to *GEO* and the results returned by *GEO* are subjected to *irrelevance checks* and then categorized according to their relevance to the disease-topic of the user-provided keywords. Users can make use of the categorized results to extract relevant results for investigation and to redirect further inspection efforts (if required). The reasons for irrelevance can also be leveraged to improve annotations of the metadata. Our experimental study demonstrates the effectiveness and superiority of ArcheGEO in identifying relevant and irrelevant matches.

The rest of the paper is organized as follows. We elaborate on the challenges to realize ArcheGEO in Section 2. We formally introduce the IMF problem in Section 3 and describe *relevant* and *irrelevant* matches in query results in Section 4. The ArcheGEO framework is detailed in Section 5. In Section 6 we report the performance of ArcheGEO. We survey related research in Section 7. The last section concludes the paper.

2 CHALLENGES

Named Entity Recognition. In the *GEO Browser*, a query is a list of keywords (*e.g.,* "breast cancer" and "human") that is provided by a user to describe their search intent whereas a returned document/result can consist of several fields (*e.g.,* `title`, `summary`, `platform`, and `organism`) in the gene expression dataset (GDS) record metadata. Identification of entities and terms (*i.e.,* named entity recognition (NER)) is usually performed on the document and query and these are then evaluated to determine if a document is *relevant* to the query. NER for biomedical terms is extremely challenging due to the technical terminology they contain and the presence of long and complex noun phrases [35]. For instance, the prostate cancer cell line MR49F has synonyms ENRZ 49F and 49FENZR which are lexically dissimilar. In addition, biomedical terms can be abbreviated, contain common English words (*e.g.,* breast adenocarcinoma in GDS5026) and may even be context dependent (*e.g.,* Cdc2 refers to two completely unrelated genes in budding and fission yeast) [35]. Further, some biomedical terms have to be inferred from the metadata. For example, in the GDS record GDS3580, we can extract the disease concept of "Pulmonary sarcoidosis" from the title. However, there is no explicit mention of lung. Instead, it has to be inferred

from biomedical ontology. Hence, generic parsers tend to perform poorly on biomedical-related queries and documents. Commercial search engines may detect some synonyms but are not sufficiently documented for reproducible pipelines or systematic reviews.

Although there are recent efforts [22, 39] toward biomedical NER using deep learning, existing biomedical NER models continue to suffer from the aforementioned limitations as the biomedical literature used for training these models contain a lot of redundant information. These redundant information can eclipse important information in the training literature yielding poor quality models [49]. Similarly, models may fail to identify named entities if the training dataset does not sufficiently cover those entities. For instance, performing disease NER using ScispaCy tool (https://scispacy.apps.allenai.org/) with *en_ner_bc5cdr_md* NER model is not able to correctly identify disease terms of the titles of GDS6063 and GDS4067. It is trained on the BC5CDR corpus (biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/) and fails to pick up "Influenza A" in GDS6063. In GDS4067, the tool picks up only "breast cancer" when in fact, two different subtypes (*i.e.,* "estrogen receptor-negative breast cancer" and "estrogen receptor-positive breast cancer") of breast cancer are mentioned. As we shall see later, ArcheGEO addresses this limitation by exploiting multiple NER models to extract different disease-related terminologies. These terminologies, which provide different facets to the concept of disease, are then leveraged for verification of the search results.

Semantic Similarity Validation. Even when biomedical-specific parsers (*e.g.,* ScispaCy [50]) are used and biomedical entities can be successfully extracted, it can still be challenging to determine if a given document (*i.e.,* result) is *relevant* to a query. This is because there is unlikely to be an exact match between the parsed biomedical entities, and the imperfect matches increase the difficulty of performing *semantic similarity* checks on concepts.

Although the creation of controlled vocabularies (*e.g., Cellosaurus* [2], *NCIt* [8]) has alleviated difficulties of evaluating *semantic similarity* of concepts, they are effective only if the vocabularies are sufficiently exhaustive, properly cross-linked internally (resp. externally) to concepts within itself (resp. contain in other vocabularies) and its usage is enforced. Many of these assumptions are violated in typical usage. Some ontologies are also lacking in contents (*e.g.,* synonyms and terms) [44]. The *Unified Medical Language System* (UMLS) represents the most extensive integration of biomedical controlled vocabularies and have been used extensively [12]. One such resource is the *UMLS Metathesaurus*, a collection of manually curated biomedical terms which have been (semi-)manually aligned [67]. Studies [46, 47] have found the manual alignments to be of high quality but incomplete.

Efforts focusing on ontology alignment research [36, 51, 53, 67] help to alleviate some of these challenges and improve cross-domain reuse and re-purposing of data. However, unresolved research challenges (*e.g.,* propagation of error in ontology models and influence of unidirectional synonyms on semantic precision and recall) remain and continue to affect the quality of ontology alignment [36]. ArcheGEO leverages disease-related concepts that span multiple ontologies and vocabularies. Multiple ontology alignments are generated as a result and this reduces the probability of missing equivalent concepts due to misaligned ontologies, thereby improving the result quality.

3 THE IMF PROBLEM

In this section, we formally describe the *irrelevance match finding* (IMF) problem addressed by ArcheGEO. We begin by introducing relevant concepts to facilitate understanding of this paper.

3.1 Terminology

***GEO* Data.** *GEO* contains *submitter-supplied* and *curated records* [14]. The submitter-supplied record is supplied by a submitter and usually summarizes an experiment. It consists of three categories of record, namely, *platform*, *sample* and *series*. A *platform* record summarizes the description of the array or sequencer and is assigned a *GEO* accession number of the form *GPLxxx*. A *sample* record describes the experimental condition of that particular sample and abundance measurement of each element in the sample. Each sample references only one platform record and is assigned a accession number *GSMxxx*. A *series* record has an accession number *GSExxx* and is a collection of related samples.

Curated records are curated from the submitter-supplied records and consist of two categories, namely, *dataset* (*i.e.,* GDS) and *profile records*. In *GEO*, a *dataset* represents a curated collection of biologically and statistically comparable *GEO samples*. Note that *samples* within the GDS are categorized into subsets based on the experimental conditions (*e.g.,* tissue or treatment) to aid comparative analysis. In addition, samples within a GDS refer to the same platform. *Profile* consists of the expression measurements of an individual gene across all samples in the GDS.

ArcheGEO works with the GDS records which comprises metadata and raw gene expression data encoded in SOFT format. The metadata consists of several default fields such as `title`, `summary`, `organism`, and `platform`. ArcheGEO exploits `title`, `summary`, and `organism` fields for *GEO* result validation. Note that GDS records are associated with experiments on three broad categories of conditions, namely, physiological disorder (*e.g.,* prostate cancer), non-disease related physiological condition (*e.g.,* ageing) and environmental condition (*e.g.,* smoking). In this paper, we focus on GDS related to physiological disorders (*i.e.,* disease) as it offers the most direct benefit to medical research. Note that ArcheGEO is extensible and can be easily configured to work for other types of records by specifying relevant concepts and features.

Disease Concept. A disease is defined as the sum of the abnormal phenomena displayed by a group of living organisms in association with a specified common characteristic by which they differ from the norm for their species in such a way as to place them at a biological disadvantage [18]. Hence, the idea of disease (*i.e., disease concept*) relates to an organism since disease-related `RNA` comes from cells and cells belong to (or were shed by) an organism. A disease is generally characterised by its associated symptoms, etiology and affected body system or tissue type. For example, β-`thalassemia` (disease concept) in human (organism) may result in anemia (symptom) and is caused by mutations in the `hemoglobin` molecule [52] (etiology). It affects the hematopoietic and lymphatic system (affected body system). Note that different synonyms can refer to the same disease concept. In *NCI Thesaurus*, both "Beta Thalassemia" and "Thalassemia Major" refer to β-thalassemia.

Controlled Vocabularies. Controlled vocabularies such as ontologies and thesauruses help to facilitate disease concept recognition. In particular, ArcheGEO exploits the *NCI Thesaurus (NCIt)* [8],

NCI Metathesaurus (NCIm) [7], *UMLS Metathesaurus* [11] and *Cellosaurus* [2] to determine equivalence of concepts extracted from user keywords and *GEO* dataset metadata.

Semantic Similarity. In general, two concepts are *equivalent* if they are *semantically similar* [45]. That is, the concepts share the same meaning. There are three broad categories of approaches to determine semantic similarity, namely, *edge-counting*-based [70], *information content (IC)*-based [56, 60] and *features*-based [57, 59] measures.

Edge-counting approaches leverage the topology of an ontology and consider shortest paths between concept pairs as a measure of semantic similarity. Although these approaches are computationally inexpensive, they suffer from several drawbacks such as the assumption that all links in the ontology represent a uniform distance, and rely on existence of cross links between ontologies for concepts found in different ontologies. In contrast, *IC*-based approaches determine semantic similarity by complementing ontology topology with information distribution of concepts in the corpus. They rely on proper disambiguation and annotation of concepts in the corpus for accurate computation of concept probability. It tends to be computationally expensive since re-computation is required whenever the ontology or corpus changes. *Features*-based approaches consider degree of overlapping between concepts as a function of their features (*i.e.,* properties) and is generally based on the Tversky's model of similarity [64] which proposes the principle that common features increase similarity while non-common ones decrease it. Due to the consideration of concept features instead of ontology topology, these approaches are more flexible and are often used to determine semantic similarity of concepts belonging to different ontologies, such as in the case of ArcheGEO.

3.2 Problem Statement

We now formally introduce the IMF problem addressed in this paper. We begin by formally defining the notion of *semantically similar*. We then use it to define the notion of *irrelevance* in the query results of *Geo Browser*. In the next section, we shall describe how to identify them.

Given two concepts A and B having features $F(A)$ and $F(B)$, respectively, the *semantic similarity* between A and B is defined as $sim(A, B) = \frac{|F(A) \bigcap F(B)|}{|F(A) \bigcup F(B)|}$. The concept A is considered to be equivalent to B if $sim(A, B) = 1$. Given two concepts A and B, and a *similarity threshold* t, A is *semantically similar* to B if $sim(A, B) \geq t$. In ArcheGEO, we consider semantic similarity between a *surrogate* (*e.g.,* metadata of returned GDS records) and a query (*e.g.,* list of user keywords) with respect to a topic (*e.g.,* "human" and "breast cancer"). A result of a query is *relevant* if its surrogate is semantically similar to the query. Otherwise, it is considered *irrelevant*.

*Definition 3.1. Given a query with concept A, a similarity threshold t, and a set of documents (*i.e., *results) D, the* ***irrelevance match finding (IMF) problem*** *identifies a set of irrelevant search results and corresponding reasons $I = \{(I_1, E_1), (I_2, E_2), \ldots, (I_n, E_n)\}$ where $I \subseteq D$, and every surrogate S_j of search result $I_j \in I$ has $sim(A, concept(S_j)) < t$ where $concept(S_j)$ yields the concept related to S_j and E_j is the reason for irrelevance of I_j.*

Observe that by identifying irrelevant results, ArcheGEO aims to improve the relevance of the search results of the *GEO Browser*.

4 RELEVANT AND IRRELEVANT MATCH

In this section, we describe how *relevance* and *irrelevance* of a query result is determined in our framework. Recall that the features of disease concept considered in ArcheGEO consist of the organism of interest, the disease synonyms (which describes the disease), anatomy of the affected body system and associated cell line. Concepts related to the first three features can be found in *NCIt, NCIm*, and *UMLS* [11]. We use *NCIt* as the base ontology as it is sufficiently extensive and cross-links with *Cellosaurus* [2] which captures cell line concepts. Amongst these features, we consider organism of interest and disease synonyms to be more important in differentiating two disease concepts as they form the fundamental definition of disease [18]. Additional features such as anatomy and cell line provide details to further disambiguate the disease concept. Although weighted-feature-based semantic similarity approach seems ideal to handle situation where features are of different importance, it requires knowledge of the feature weight which is ambiguous in this case. Instead, we adopt a *rule-based approach*. Note that a pair of concepts that is neither semantically relevant nor irrelevant is considered as *ambiguous*.

Since the GDS record metadata provides an `organism` field by default, we assume that it is always possible to conclude if the organism feature between the metadata and user keywords show a match or a mismatch. A keen reader may observe that since we do not impose constraints on query keywords, it is possible for a user to exclude keywords associated to organism of interest in her query. In this case, the query is assumed to imply no specific requirement regarding organism feature. That is, the disease feature semantic check always return an *equivalent* relationship.

Definition 4.1. Let $O(D)$, $S(D)$, $A(D)$ and $C(D)$ be the organism, synonym, anatomy and cell line feature of disease D, respectively. Given two diseases D_1 and D_2, D_1 is ***semantically relevant*** *to D_2 if any of the following condition is satisfied:*

- $O(D_1) \Leftrightarrow O(D_2)$ *and* $S(D_1) \Leftrightarrow S(D_2)$
- $O(D_1) \Leftrightarrow O(D_2)$ *and* $S(D_1) \leftrightsquigarrow S(D_2)$ *and* $(A(D_1) \Leftrightarrow A(D_2)$ *or* $C(D_1) \Leftrightarrow C(D_2))$

where $X \Leftrightarrow Y$ and $X \leftrightsquigarrow Y$ represent equivalence and ambiguous relationship, respectively, between X and Y.

We consider two features to be *equivalent* if they are annotated by the same identifier in a given controlled vocabulary. An *ambiguous* relationship can occur if at least one feature (*e.g.,* $C(D_1)$ in the feature pair (*e.g.,* $C(D_1)$ and $C(D_2)$) being compared is null (*e.g.,* there is no cell line information associated with D_1). Note that it is possible for a feature X to be multi-valued. In this case, feature X (*e.g.,* $X = \{breast\ cancer, diabetes\}$) is equivalent to feature Y (*e.g.,* $Y = \{breast\ cancer, breast\ carcinoma\}$) if X and Y have at least one common value (*e.g.,* $X \bigcap Y = \{breast\ cancer\}$).

Definition 4.2. Following from Def. 4.1, D_1 is ***semantically irrelevant*** *to D_2 if any of the following condition is satisfied:*

- $O(D_1) \not\Leftrightarrow O(D_2)$
- $S(D_1) \not\Leftrightarrow S(D_2)$
- $S(D_1) \leftrightsquigarrow S(D_2)$ and $(A(D_1) \not\Leftrightarrow A(D_2)$ or $C(D_1) \not\Leftrightarrow C(D_2))$

Observe that Definition 4.1 is based on the concept of semantically similar (Section 3.2). In particular, the threshold $t = 1$ in

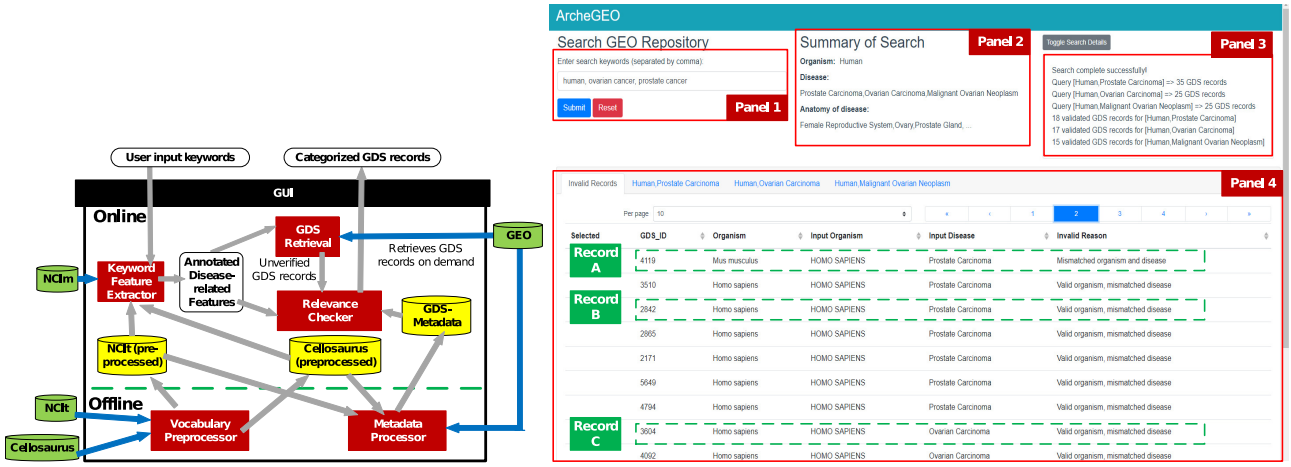

Figure 2: Architecture (left) and GUI (right) of ArcheGEO.

Definition 4.1. In addition, it introduces additional conditions (*i.e.,* ambiguous relationships) for semantic similarity.

5 THE ARCHEGEO FRAMEWORK

Figure 2 (left) depicts the architecture of ArcheGEO. It consists of two major components. The *offline* component handles the preprocessing of controlled vocabularies and GDS metadata, whereas the *online* component performs relevance validation. We elaborate on them in turn. Proofs of all lemmas are given in [23].

ArchGEO is provided as a web service and the **code** is available at https://github.com/ArcheGEO/hogwarts-master. An end user may access it using the GUI in Figure 2 (right). *Panel 1* takes in the user keywords (separated by comma) as input. In the online component, the *Keyword Feature Extractor* extracts disease concept-related features from the keywords and annotates them using knowledge stores that are generated by the offline component. The summarized features are presented in *Panel 2*. Then, *GDS Retrieval* leverages the features to reconstruct a set of appropriate queries that is sent to *GEO*. The GDS records returned by *GEO* are subjected to semantic similarity checking in the *Relevance Checker*. These records are then categorized based on the relevance/irrelevance of the extracted features from the metadata to that of the user keywords. *Panel 3* shows a summary of the search results. A browsable list of the categorized results is presented in *Panel 4*. Note that both relevant and irrelevant categories of matches are displayed in *Panel 4*.

In particular, ArcheGEO provides reasons for the irrelevance (Def. 3.1) which can be leveraged for correcting misannotations. For example, *Record C* in Figure 2 indicates a "disease mismatch" of GDS3604 for the keyword "ovarian cancer", highlighting a potential misannotation of GDS3604. A review of the GDS title (*i.e.,* "Tamoxifen effect on endometrioid carcinomas") can be conducted to extract relevant disease-related terms (*e.g.,* "endometrioid carcinomas") to improve the quality of record annotation.

5.1 The Offline Component

ArcheGEO requires specific features (*i.e.,* organism, synonyms, term identifier, anatomy and cell line) from the controlled vocabularies and the GDS metadata to perform relevance validation. These features can either be extracted from various repositories on demand (*i.e.,* online) or can be preprocessed (*i.e.,* offline) to reduce wait time. We choose the latter approach to improve usability of ArcheGEO. The offline component consists of a *Vocabulary Preprocessor* and a *Metadata Preprocessor* that extracts required features from controlled vocabularies and GDS metadata, respectively. The extracted features are stored as an internal knowledge base using a PostgreSQL database for subsequent usage.

Vocabulary Retrieval. We preprocess *NCIt* and *Cellosaurus* and store the extracted features. In particular, features such as concept identifier and synonyms of anatomy, organism and disease/abnormality are extracted from the *Thesaurus.OWL* (https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/) by leveraging inherent XML tags (*e.g.,* <P90> and <P331> refer to synonym and *NCBI Taxon ID*, respectively). Similarly, concept identifier, synonyms of cell line and *NCIt* identifier of diseases associated with particular cell line are extracted from *Cellosaurus* (Cellosaurus.txt, https://ftp.expasy.org/databases/cellosaurus) using regular expression that match tags such as SY which symbolizes synonyms.

Metadata Extraction. Figure 3 depicts the metadata extraction process realized by the *Metadata Preprocessor*. We utilize the `organism` field to extract the organism feature, and `title` and `summary` fields to extract synonyms, anatomy and cell line features. We store the title-derived features separately from the summary-derived features, as they provide different granularity of information. The SOFT file FTP links of the GDS records are also extracted to facilitate download of raw gene expression data.

Named entity recognition (NER) is performed on `title` and `summary` fields using ScispaCy [50], a specialized NLP library for processing biomedical texts (*i.e.,* bioNLP). Note that the choice of a suitable bioNLP is orthogonal to the IMF problem addressed by ArcheGEO. We choose ScispaCy over other bioNLPs such as *MetaMap* [13] and *MetaMapLite* [25] as it is more efficient [50]. ArcheGEO utilizes the BC5CDR (*en_ner_bc5cdr_md*) and JNLPBA (*en_ner_jnlpba_md*) NER models provided by ScispaCy to extract features of disease synonyms and cell lines, respectively. The organism feature is extracted from the GDS metadata directly.

Once the features are identified and extracted, they are annotated with *NCIt* and *Cellosaurus* identifiers (where appropriate) to facilitate relevance validation in the online component. For disease

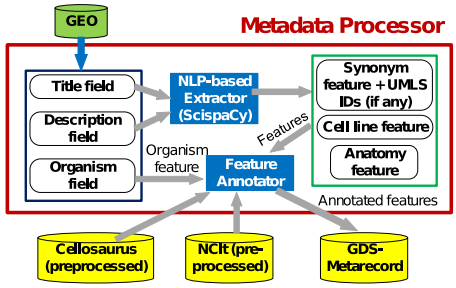

Figure 3: Metadata preprocessor.

synonym feature, ArcheGEO leverages the ScispaCy's entity linker to extract associated *UMLS* identifiers if they are available. There is linkage between records in *NCIt* and those in *UMLS* facilitating annotation (with *NCIt* identifiers) of these features. Organism features are mapped to *NCIt* entries by looking for exact matches between the feature and synonyms of the entries. In the case of cell line features, mapping between extracted NER and *Cellosaurus* entry is less straight forward. The extracted NER tends to contain additional English words whereas synonyms in *Cellosaurus* are mostly scientific names. For example, parsing the title of GDS4121 with ScispaCy yields "prostate cancer DU145 cell line" for cell line-type entity whereas known synonyms in *Cellosaurus* are {DU-145, DU145, DU 145, DU_145, DU.145, Duke University 15}. A naive string match algorithm that seeks to find an exact match of "prostate cancer DU145 cell line" within the synonym list will fail to return a match. A more effective approach would be to find the existence of synonyms within the extracted entity. Briefly, we tokenize the cell line feature and performs matching of each token with the synonym. The cell line feature is mapped to a *Cellosaurus* entry if at least one token matches at least one synonym.

Unlike organism, synonyms and cell line features which are usually explicitly stated in the metadata, the annotation of the anatomy is usually inferred from annotation of synonyms and cell line features. It infers anatomy annotation by using ontological relationships in *NCIt* (*i.e., Disease_Has_Primary_Anato-mic_Site* and *Disease_Has_Associated_Anatomic_Site*) and *Cellosaurus* (*i.e., Derived from metastatic site* and *Derived from sampling site*). It is possible for a feature that has valid correspondence to *NCIt* or *Cellosaurus* entries to lack appropriate annotation due to the ambiguous nature of natural text. Since missed annotation impacts relevance checking, it associates the disease concept with multiple features and determines relevance based on multiple features.

Lemma 5.1. *The worst-case time complexity to process metadata of GDS (denoted as M_{GDS}) is $O(|M_{GDS}|(n^2 + p \times k))$ where n is the maximum number of words in the title and summary fields of GDS metadata and k (resp. p) is the maximum number of values (resp. tokens) associated to a feature (resp. the cell line feature).*

5.2 The Online Component

The online component consists of three key subcomponents, namely, *Keyword Feature Extractor*, *GDS Retrieval* and *Relevance Checker*.

Feature Extraction from Keywords. The *Keyword Feature Extractor* extracts features that are related to the disease concept from the user input keywords (*e.g.,* "human", "breast cancer"). A particular keyword is linked to a specific feature if it is found in

Algorithm 1 GDS Retrieval.

Require: Set of organism keywords K_O, set of synonym keywords K_S;
Ensure: Set of unverified GDS records R_{GDS};
1: $initialize(R_{GDS})$
2: $Q \leftarrow createSQLQueries(K_O, K_S)$
3: **for** each $q \in Q$ **do**
4: $ID_{GDS} \leftarrow eSearch(q)$ /*pipes q to *eSearch* utility */
5: **for** each $i \in ID_{GDS}$ **do**
6: $R_{GDS} \leftarrow R_{GDS} \bigcup \{downloadAndDecompressFile(i)\}$
7: **end for**
8: **end for**

the set of synonyms associated to that feature. For example, the user keyword "human" is associated to the organism feature "homo sapiens" (*NCIt ID* = C14225) since "human" is contained in the set of synonyms {"Homo Sapiens", "Human", "Human, General"} that is associated with the concept C14225 in *NCIt*. It annotates the keyword with corresponding *NCIt* identifiers where appropriate.

GDS Retrieval. ArchGEO does not assume that users would leverage logical operators such as *OR* and *AND* to formulate queries. Instead, the only requirement on keywords is that they should be comma-delimited. Algorithm 1 uses the annotated disease-related keywords to send a batch of queries to *GEO*. In particular, the disease synonym keywords (K_S) and organism keywords (K_O) are used to construct all possible disease-organism pair-wise queries (*i.e.,* rewritten queries) (Line 2). For example, in Figure 2(b), the rewritten queries are: (1) "human" *AND* "prostate cancer" and (2) "human" *AND* "ovarian cancer". Note that by organizing the queries as such, one can easily obtain the desired result by recombining these results through ArchGEO. For instance for the above query, one can easily view results of either "prostate cancer" *AND* "ovarian cancer" or "prostate cancer" *OR* "ovarian cancer" in human by indicating their preference through the GUI. In ArchGEO GUI, the results of each rewritten query are presented in a tab in *Panel 4* (Figure 2). *GDS Retrieval* pipes the rewritten queries to *GEO* using the *eSearch* program of the programmatic access utility (Lines 3-8). Briefly, *eSearch* returns a list of unique identifiers (*i.e.,* GDS identifier) that matches a given text query (Line 4). We refer to these as *unverified GDS records*. GDS record files matching these identifiers are then downloaded and decompressed.

Relevance and Irrelevance Matching. Algorithm 2 outlines the procedure of the *Relevance Checker* subcomponent which is responsible for finding relevant and irrelevant matches in the query results. For each unverified GDS record, it retrieves disease-related concept features of its metadata from GDS-Metadata store (Lines 3-7) and performs semantic relevance check (Lines 8 and 10) against the features extracted from the user keywords based on Definitions 4.1 and 4.2. The importance of a feature can be influenced by its occurrence in a specific structure (*i.e.,* title, summary or abstract, main body) of a document [62]. Intuitively, the level of importance decreases as we move from title to summary to main body since additional description is added and it not only serves to enrich the topic of interest, but may also add additional noise [43] which impacts checks on semantic similarity. Hence, we adopt a *granulated validation* approach. Semantic similarity of features derived from the `title` field (Line 8) is considered to be more significant than that of features derived from the `summary` field (Line 10). In particular, we first check for relevance (and irrelevance) of GDS records with respect to the query using features from the `title` field. When the

Algorithm 2 Relevance Check.

Require: Set of metadata of unverified GDS records M_{GDS}, set of organism keywords K_O and synonym keywords K_S;
Ensure: Relevance of GDS records REL_{GDS};
1: $initialize(REL_{GDS})$
2: **for** each $m \in M_{GDS}$ **do**
3: $m_t \leftarrow getTitleField(m)$
4: $m_s \leftarrow getSummaryField(m)$
5: $O(m) \leftarrow getOrganismFeature(m)$
6: $S(m_t), A(m_t), C(m_t) \leftarrow getOtherFeatures(m_t)$
7: $S(m_s), A(m_s), C(m_s) \leftarrow getOtherFeatures(m_s)$
8: $REL_m \leftarrow checkRelevance(O(m), S(m_t), A(m_t), C(m_t), K_O, K_S)$
9: **if** $isAmbiguous(REL_m) == true$ **then**
10: $REL_m \leftarrow checkRelevance(O(m), S(m_s), A(m_s), C(m_s), K_O, K_S)$
11: **end if**
12: $REL_{GDS} \leftarrow REL_{GDS} \bigcup \{REL_m\}$
13: **end for**

record is deemed to be ambiguous (Line 9), we proceed to validate relevance using features from the summary field. As we shall see in Section 6.2, this granulated validation approach achieves good precision and recall compared to alternative strategies.

A record is classified as *valid* or *invalid* based on the aforementioned check. *Valid* records are those that are semantically relevant. We also consider a record valid if the organism features of the metadata and user keywords are equivalent but relationships of remaining features are ambiguous. However, since the validity of such a record is less certain, an explanation "Valid Organism, Uncertain Disease" is attached with it to allow a user to further verify their validity through other means such as literature review. *Invalid* records are those that are found to be semantically irrelevant. The exact reason (*i.e.,* explanation) for semantic irrelevance is generated as follows . Given a record I_j, if $O(D_1) \nLeftrightarrow O(D_2)$ then E_j ="Organism Mismatched" (*e.g.,* Record A in Figure 2); if $S(D_1) \nLeftrightarrow S(D_2)$ or $S(D_1) \leftrightsquigarrow S(D_2)$ and ($A(D_1) \nLeftrightarrow A(D_2)$ or $C(D_1) \nLeftrightarrow C(D_2)$) then E_j = "Disease Mismatched" (*e.g.,* Record B in Figure 2) where D_1 and D_2 refer to the disease concepts obtained from the GDS metadata and user keywords, respectively.

Lemma 5.2. *The worst-case time and space complexities to perform relevance checks are $O(k|R_{GDS}|)$ and $O(k|R_{GDS}|)$, respectively, where k is the maximum number of distinct values associated to any feature and R_{GDS} is the set of unverified GDS records.*

6 PERFORMANCE STUDY

The online component of ArcheGEO is implemented as a web service. The front-end and back-end are implemented using *Vue* and *Spring Boot*, respectively. The offline component uses *Python 3.8* and the ScispaCy library for NLP processing and feature extraction. PostgreSQL is used to store the preprocessed vocabularies in ArcheGEO. In this section, we investigate the performance of ArcheGEO and report the key findings. A case study is discussed in [23]. All experiments are performed on a 64-bit Windows desktop with Intel(R) Core(TM) i7-4790K CPU (4GHz) and 32GB of main memory.

6.1 Experimental Setup

Two sets of experiments are carried out. The first set (*Exp 1* and *2*) examines design decisions affecting performance of ArcheGEO whereas the second set (*Exp 3* and *4*) examines ArcheGEO's performance against benchmark systems and its usefulness.

Table 1: Test Collection.

TC	Query Keywords	Total Records	# Of Relevant
TC1	{ *human, breast cancer* }	173	133
TC2	{ *human, type 2 diabetes mellitus* }	15	7
TC3	{ *human, ovarian carcinoma* }	25	18
TC4	{ *human, prostate carcinoma* }	35	22
TC5	{ *human, endometriosis* }	8	7
TC6	{ *human, parkinson disease* }	21	7
TC7	{ *human, malaria* }	11	6
TC8	{ *human, psoriasis* }	25	11
TC9	{ *mus musculus, parkinson disease* }	10	7
TC10	{ *mus musculus, breast cancer* }	42	20
TC11	{ *mus musculus, ovarian cancer* }	9	3
TC12	{ *mus musculus, type 2 diabetes mellitus* }	10	8
TC13	{ *plasmodium falciparum, malaria* }	11	3
TC14	{ *rattus norvegicus, type 2 diabetes mellitus* }	4	3
TC15	{ *rattus norvegicus, heart disease* }	20	17
TC16	{ *rattus norvegicus, pulmonary disease* }	12	12
TC17	{ *mus musculus, arthritis* }	11	8
TC18	{ *mus musculus, lung cancer* }	12	6
TC19	{ *rattus norvegicus, parkinson disease* }	3	3
TC20	{ *rattus norvegicus, liver disease* }	13	8

Benchmarks. We are unaware of any existing *GEO* record retrieval systems that specifically identify irrelevant records from the *GEO Browser* results. Hence, we are confined to compare ArchGEO with record retrieval systems (*i.e.,* ScanGEO [37] and DataMed [20]) that can obtain *GEO* records based on user-specified keywords. ScanGEO and DataMed can be found at http://scangeo.dartmouth.edu/ScanGEO/ and https://datamed.org/, respectively.

Test Collections (TC). For our experiments, the test collections are result sets retrieved from *GEO Browser* for given sets of keyword-based queries. Five postgraduate students doing biomedical research and familiar with the *GEO* datasets volunteered to perform relevance judgement on the test collections (Table 1). Every GDS record in each collection is assigned as either being relevant or irrelevant to the query based on majority voting. The kappa value of record judgement varies in the range [0.68 - 1], indicating good to excellent agreement of the judgement [17].

Performance Evaluation. We follow the Cranfield paradigm [65] for evaluating ArcheGEO. In particular, domain experts provide judgements regarding relevance of topical similarity on test collections. All systems are then assessed based on precision and recall on the test collections. In all experiments, the similarity threshold t of ArcheGEO is set to 1.

6.2 Experimental Results

Exp 1: Metadata Structure-based Relevance Validation. We evaluate the effect of performing relevance validation based on the location (*i.e.,* title and summary fields) of features in the metadata using TC1 (Figure 4). We examine 5 strategies for considering features in relevance validation, namely, title field alone (*T*), summary field alone (*S*), title and summary fields together (*All*), title field then summary field (*TS*), summary field then title field (*ST*). In *T* and *S*, only features in the specified field are considered. For *All*, each feature is formed by performing a union of the features extracted from title and summary fields. *TS* corresponds to the approach described in Section 5.2. *ST* refers to the approach when title and summary fields are swapped in *TS*. We observed that the recall of both *S* and *ST* are poorer than that of *T* and *TS* (Figure 4, top left). This agrees with the observation from [43] that summary field is likely to be noisier than title field, thereby affecting relevance

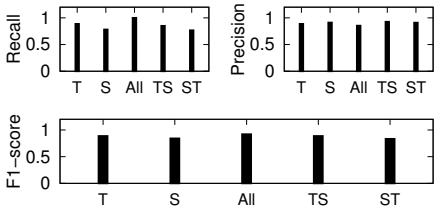

Figure 4: Effect of various relevance validation strategies.

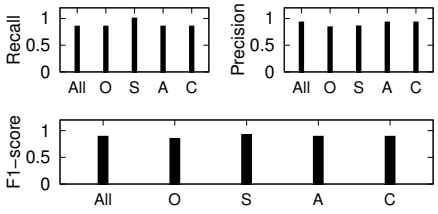

Figure 5: Effect of disease-related features.

validation. Although *All* yielded the best F1-score (*i.e.,* 0.92), its precision is also the poorest (*i.e.,* 0.85). *TS* which has the highest precision (*i.e.,* 0.93) has comparable recall and F1-score as *T*. Hence, we select *TS* as our relevance validation strategy.

Exp 2: Effect of Disease-related Features. Next, we examine the effect of the features (*i.e.,* organism (*O*), synonym (*S*), anatomy (*A*) and cell line (*C*)) on relevance validation using TC1 (Figure 5). Variants of ArcheGEO are generated by excluding each feature in turn from the validation. For clarity, *X* (*i.e.,* variant-*X*) in Figure 5 refers to the variant of ArcheGEO where feature *X* is excluded and *All* refers to ArcheGEO. We observed that the exclusion of features *O* (0.84) and *S* (0.85) affected the precision of ArcheGEO (Figure 5, top right). In particular, when features *O* and *S* are excluded, an increase in the number of ambiguous records that are classified as valid is observed. All valid records of variant-*O* were categorized as *"Organism Unverified"* and within these records, an additional 7 were categorized as *"Disease Unverified"* as well. In the case of variant-*S*, 30 records were labelled as *"Disease Unverified"*. This number reduced to five for variant-*A* and variant-*C*. *This highlights the importance of features O and S in comparison to A and C, justifying our choice of these features as the dominant discriminating factors between two diseases (Section 4).* Based on Figure 5, the inclusion of features *A* and *C* may seem redundant. However, these features are still useful for relevance validation, especially in the case when feature *S* is absent from the metadata. For instance, in the case of variant-*S*, 103 out of 133 valid records were categorized as *"Disease Valid"*. That is, based on features *A* and *C* alone, 77.4% of valid records are still correctly classified as *"Disease Valid"*.

Exp 3: Comparison with Benchmark Systems. Next, we examine the recall, precision, and F1-score of ArcheGEO, ScanGEO and DataMed on TC1 to TC20. Figure 6 reports representative results. In particular, DataMed (resp. ScanGEO) did not retrieve any results for TC2 to TC7, TC16 and TC19 (resp. TC6, TC9, TC16 and TC19). The recall for ScanGEO, DataMed and ArcheGEO vary in the range of [0 - 0.91], [0 - 1] and [0.18 - 1], respectively while that of precision are in the range of [0 - 1], [0 - 1] and [0.5 - 1], respectively. Although ScanGEO (resp. DataMed) performs better (up to 1.5X) in terms of precision, its recall is up to 12X (resp. 28.3X) poorer compared to ArcheGEO for the test collections. In particular, the F1-score of ArcheGEO is up to 6.9X (resp. 15.7X) better than ScanGEO (resp. DataMed). Further, the range of recall and precision is tighter for ArcheGEO compared to both ScanGEO and DataMed, highlighting more consistent performance.

In addition, we measure the wall-clock time taken for each query. For clarity, the time duration is from the instance the search is invoked (*i.e.,* clicking of "search" button) until the desired result set is retrieved. ArcheGEO performs moderately in terms of runtime and is able to complete the relevance validation and categorization within 10 seconds. DataMed is the slowest. This is likely because DataMed is an open source discovery index which references multiple sources (including *GEO*) and a user has to take an additional step to configure it to display only *GEO*-specific results. Hence, *ArcheGEO can effectively and efficiently identify relevant results from the GEO result set.*

Exp 4: Irrelevance Matches. Lastly, we characterize the valid and invalid results (Section 5.2) obtained from ArcheGEO in terms of the percentage of records that are relevant (Def. 4.1), irrelevant (Def. 4.2), and ambiguous (*i.e.,* neither relevant nor irrelevant) in Figure 7 (top). Note that in ArcheGEO GUI, ambiguous records are reflected as valid records with "Valid Organism, Uncertain Disease" tags. All cases report some irrelevant records except TC19. *On average, 44.9% of results are identified as irrelevant, highlighting a need for validating the search results.* The average percentage of relevant and ambiguous results are 41.5% and 13.7%, respectively.

We also examine the reasons for irrelevance. Figure 7 (bottom) reports the percentage of records that fall under different categories, namely, mismatched organism (*i.e., "Organism Mismatch, Disease Valid"*, *"Organism Mismatch, Disease Unverified"*), mismatched disease (*i.e., "Organism Valid, Disease Mismatch"*) and mismatched organism & disease (*i.e., "Organism & Disease Mismatch"*). On average, 25.8%, 49.5% and 19.7% of records reported to be irrelevant are due to mismatched organism, mismatched disease, and both mismatch organism & disease, respectively. Interestingly, even though GDS metadata contain a dedicated `organism` field, we observe that 45.5% of returned records (*i.e.,* mismatch organism and mismatch organism & disease) contain a mismatch between its metadata organism information and the actual organism requested by users. Enriching the GDS metadata with additional fields such as disease and exploiting these data can further improve the quality of the search results.

7 RELATED WORK

***GEO*-related Software**. Several software tools have been proposed to support the usage of *GEO*. They can be broadly classified into four categories: (1) data conversion (*e.g.,* GEOquery [24], GEOmetadb [73]), (2) data analysis (*e.g.,* shinyGEO [28], ScanGEO [37], D-GEX [21]), (3) records retrieval (*e.g.,* DataMed [20], ScanGEO [37]) and (4) metadata curation (*e.g.,* GEOMetaCuration [42], crowdsourced curation [69]). GEOquery is focussed on converting *GEO* data into a format that is compatible with BioConductor whereas GEOmetadb is a MySQL implementation for storing *GEO* metadata for efficient querying and retrieval of *GEO* metadata. Efforts such as [42] and [69] aim to increase the ease of annotation and curation of metadata. D-GEX uses deep learning-based approach to infer the expression of target genes from the expression of landmark genes.

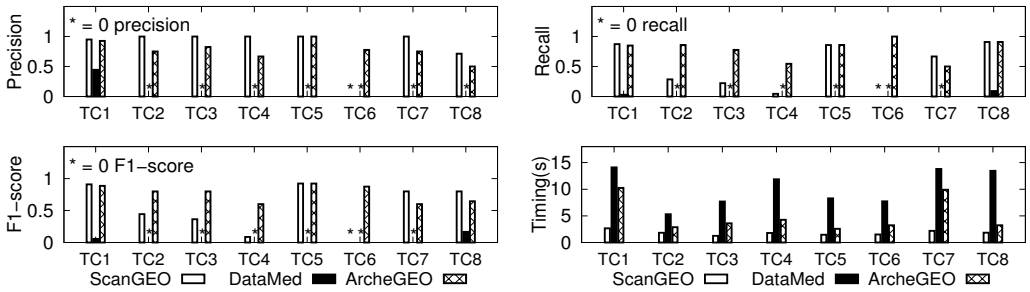

Figure 6: Comparison with benchmark systems.

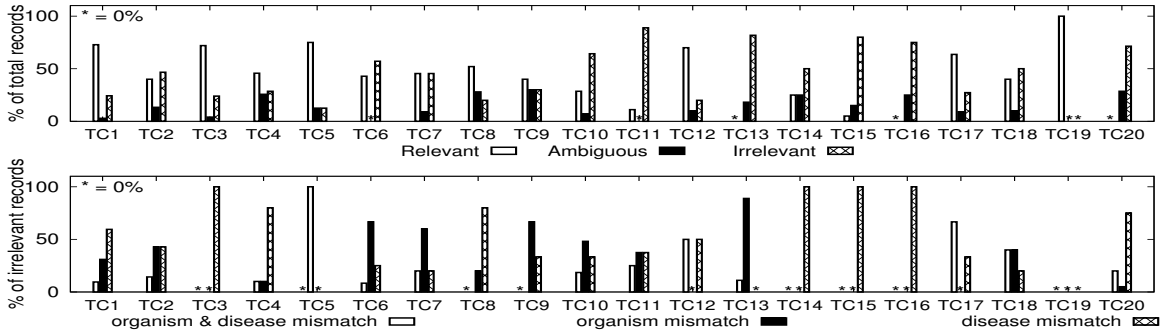

Figure 7: Characteristics of ArcheGEO results.

shinyGEO and ScanGEO are used for performing differential expression analysis. In addition, ScanGEO provides a keyword-based *GEO* record retrieval browser interface. DataMed in contrast, is an open source discovery index for finding biomedical datasets that includes *GEO*. In contrast, ArcheGEO is designed to work as a companion of *GEO* query interface and its goal is to address the IMF problem (*i.e., improve* search result quality by locating irrelevant records). A key difference between ArcheGEO and existing *GEO* record retrieval system is its ability to report on irrelevant records and their reasons of irrelevance.

Relevance Feedback. Relevance feedback is the process of obtaining user feedback regarding relevance of documents. Typically, a search engine will present a set of retrieved documents to the users for them to indicate whether they are relevant. Based on the feedback, it modifies the query to retrieve a new set of results. ArcheGEO differs from such systems (*e.g.,* [29, 58]) as the main goal is to predict and categorize the relevance of the search results for presentation to users instead of seeking user's feedback regarding search relevance, which is a time-consuming task in *GEO*.

Search Result Organization. The most common form of organization is the ranked list (*i.e.,* relevance ranking) where results are ordered according to their probability of being relevant to the user's query. Examples include DeepRank [54] and ranking in Yahoo search [71]. The effectiveness of the ranked list relies on the user being able to input appropriate query for the desired documents. The use of less relevant keywords may result in searching through a long ranked list before finding desired documents [61]. An alternative approach is document clustering which groups similar results together. Ranking-based approaches are more suited for regression-type problems whereas clustering-based approaches (*e.g.,* [40], [61]) are more appropriate for classification-type problems such as that in ArcheGEO where the goal is to distinguish relevant results from irrelevant ones. In particular, ArcheGEO categorizes the results into two broad classes (relevant and irrelevant results). The relevant results are further organized based on disease-organism categories whereas irrelevant results are further classified according to their reasons of irrelevance. Unlike [40], the relevant categories are not ranked as the disease-organism categories are derived from the query in which all keywords are assumed to be of equal importance. In [61], the research goal is focused on how to improve the quality of cluster labels. This is different from ArcheGEO which aims to improve result quality through classification of search results into relevant and irrelevant clusters.

8 CONCLUSIONS

In this paper, we describe an end-to-end framework called ArcheGEO which is targeted towards users of the *GEO* repository, and delivers value by separating relevant and irrelevant matches in query results, thereby improving access to relevant biomedical information. Reasons for irrelevance are reported and such details can be used to correct misannotations or to include missing annotations of GDS records in *GEO* repository to improve search results relevance.

Several open challenges still await in this space. First, the framework can be expanded to other data in *GEO* (*e.g.,* series and profiles) and to other GDS types (*e.g.,* environmental records). Additional features can also be included to improve the relevance validation process. Second, improved NER techniques can improve the overall quality of the relevance validation. Third, index optimization [26] is extremely useful for efficient keyword search that can further reduce the validation time. Lastly, it would be interested to explore how ArcheGEO could be applied to tools which are processing vast quantities of sequencing data (*e.g.,* https://rna.recount.bio/).

Acknowledgements. The authors are supported by the AcRF Tier-2 Grant MOE2019-T2-1-029.

REFERENCES

[1] ArrayExpress. https://www.ebi.ac.uk/arrayexpress/.
[2] Cellosaurus. https://web.expasy.org/cellosaurus/.
[3] Classification of Diseases. https://www.who.int/standards/classifications/classification-of-diseases.
[4] Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/.
[5] Genomic Expression Archive. https://www.ddbj.nig.ac.jp/gea/index-e.html.
[6] Medical Subject Headings. https://www.ncbi.nlm.nih.gov/mesh/.
[7] NCI Metathesaurus. https://ncim.nci.nih.gov/ncimbrowser/.
[8] NCI Thesaurus. https://ncithesaurus.nci.nih.gov/ncitbrowser/.
[9] Online Mendelian Inheritance in Man. https://www.omim.org/.
[10] SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html.
[11] UMLS Metathesaurus. https://uts.nlm.nih.gov/uts/umls/home.
[12] L. Amos, et al. UMLS users and uses: a current overview. *Journal of the American Medical Informatics Association*, 27(10): 1606-1611, 2020.
[13] A.R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. *Proc AMIA Symp*, 17-21, 2001.
[14] T. Barrett, et al. NCBI GEO: mining tens of millions of expression profiles - database and tools update. *Nucleic Acids Research*, 35(suppl_1): D760-D765, 2007.
[15] H. Bono. All of gene expression (AOE): an integrated index for public gene expression databases. *PloS one*, 15(1): e0227076, 2020.
[16] M. Brockington, et al. Localization and functional analysis of the LARGE family of glycosyltransferases: significance for muscular dystrophy. *Human Molecular Genetics*, 14(5): 657-665, 2005.
[17] T. Byrt. How good is that agreement? *Epidemiology*, 7(5): 561, 1996.
[18] E.J.M. Campbell, J.G. Scadding, R.S. Roberts. The concept of disease. *Br Med J*, 2(6193): 757-762, 1979.
[19] G. Chen,et al. Restructured GEO: restructuring gene expression omnibus metadata for genome dynamics analysis. *Database*, 2019.
[20] X. Chen, et al. DataMed - an open source discovery index for finding biomedical datasets. *Journal of the Americal Medical Informatics Association*, 25(3): 300-308, 2018.
[21] Y. Chen, et al. Gene expression inference with deep learning. *Bioinform.*, 32(12): 1832-1839, 2016.
[22] H. Cho, H. Lee. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, 20(735), 2019.
[23] H.-E Chua, L. Tucker-Kellogg, S. S. Bhowmick. ArcheGEO: Towards improving relevance of gene expression omnibus search results. Technical Report, https://personal.ntu.edu.sg/assourav/TechReports/ArcheGEO-TR.pdf, 2021.
[24] S. Davis, P.S. Meltzer. GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics*, 23(14): 1846-1847, 2007.
[25] D. Demner-Fushman, W.J. Rogers, A.R. Aronson. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J Am Med Inform Assoc*, 24(4): 841-844, 2017.
[26] B. Ding, et al. Optimizing index for taxonomy keyword search. *In SIGMOD*, 2012.
[27] D. Djordjevic, et al. Discovery of perturbation gene targets via free text metadata mining in gene expression omnibus. *Computational Biology and Chemistry*, 80: 152-158, 2019.
[28] J. Dumas, M.A. Gargano, G.M. Dancik. shinyGEO: a web-based application for analyzing gene expression omnibus datasets. *Bioinformatics*, 32(23): 3679-3681, 2016.
[29] G. Gay, et al. On the use of relevance feedback in IR-based concept location. *In IEEE ICSM*, 2009.
[30] C.B. Giles,et al. ALE: automated label extraction from GEO metadata. *BMC Bioinformatics*, 15(14): 7-16, 2017.
[31] E.S. Gushchanskaia, et al. Interplay between small RNA pathways shapes chromatin landscapes in C. elegans. *Nucleic Acids Research*, 47(11): 5603-5613, 2019.
[32] D. Hadley, et al. Precision annotation of digital samples in NCBI's gene expression omnibus. *Scientific Data*, 4(1): 1-11, 2017.
[33] A.N. Hasan, et al. An in silico analytical study of lung cancer and smokers datasets from gene expression omnibus (GEO) for prediction of differentially expressed genes. *Bioinformation*, 11(5): 229, 2015.
[34] R.Q. He, et al. Clinical significance of miR-210 and its prospective signaling pathways in non-small cell lung cancer: evidence from gene expression omnibus and the cancer genome atlas data mining with 2763 samples and validation via real-time quantitative PCR. *Cellular Physiology and Biochemistry*, 46(3): 925-952, 2018.
[35] L.J. Jensen, J. Saric, P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2): 119-129, 2006.
[36] N. Karam, et al. Matching biodiversity and ecology ontologies: challenges and evaluation results. *The Knowledge Engineering Review*, 35(E9): 1-19, 2020.
[37] K. Koeppen, B.A. Stanton, T.H. Hampton. ScanGEO: parallel mining of high-throughput gene expression data. *Bioinformatics*, 33(21): 3500-3501, 2017.
[38] Y.S. Lee, et al. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29(23): 3036-3044, 2013.
[39] J. Lee,et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234-1240, 2020.
[40] A. Leuski. Evaluating document clustering for interactive information retrieval. *In CIKM*, 2001.
[41] Y. Li, et al. SCIA: a novel gene set analysis applicable to data with different characteristics. *Frontiers in Genetics*, 10: 598, 2019.
[42] Z. Li, J. Li, P. Yu. GEOMetaCuration: a web-based application for accurate manual curation of gene expression omnibus. *Database*, 2018, 2018.
[43] J. Lin. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10(1): 1-15, 2009.
[44] S. Mathur, D. Dinakarpandian. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, 45(2): 363-371, 2012.
[45] R. Mihalcea, C. Corley, C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. *In AAAI*, 2006.
[46] C.P. Morrey, et al. Resolution of redundant semantic type assignments for organic chemicals in the UMLS. *Artificial Intelligence in Medicine*, 52(3): 141-151, 2011.
[47] F. Mougin, N. Grabar. Auditing the multiply-related concepts within the UMLS. *Journal of the American Medical Informatics Association*, 21(e2): e185-e193, 2014.
[48] C.J. Mungall,et al. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1): 1-20, 2012.
[49] U. Naseem, et al. Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding. *In IJCNN*, 2020.
[50] M. Neumann, et al. ScispaCy: fast and robust models for biomedical natural language processing. *In BioNLP*, 2019.
[51] V. Nguyen, H.Y. Yip, O. Bodenreider. Biomedical vocabulary alignment at scale in the umls metathesaurus. *In Proceedings of the Web Conference*, 2021.
[52] A.W. Nienhuis, D.G. Nathan. Pathophysiology and clinical manifestations of the β-thalassemias. *Cold Spring Harbor Perspectives in Medicine*, 2(12): a011726, 2016.
[53] D. Oliveira, C. Pesquita. Improving the interoperability of biomedical ontologies with compound alignments. *J. Biomed. Semant.*, 9(1), 2018.
[54] L. Pang, et al. Deeprank: A new deep architecture for relevance ranking in information retrieval. *In CIKM*, 2017.
[55] E.G. Puffenberger, et al. Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYL loss of function. *Proceedings of the National Academy of Sciences*, 101(32): 11689-11694, 2004.
[56] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *In IJCAI*, 1995.
[57] M.A. Rodríguez, M.J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2): 442-456, 2003.
[58] Y. Rui, et al. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5): 644-655, 1998.
[59] D. Sánchez, et al. Ontology-based semantic similarity: a new feature-based approach. *Expert Systems with Applications*, 39(9): 7718-7728, 2012.
[60] N. Seco, T. Veale, J. Hayes. An intrinsic information content metric for semantic similarity in WordNet. *In ECAI*, 2004.
[61] H. Toda, R. Kataoka. A search result clustering method using informatively named entities. *In WIDM*, 2005.
[62] A. Trotman. An artificial intelligence approach to information retrieval. *In SIGIR*, 2004.
[63] D. Tsoucas, et al. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1): 1-9, 2019.
[64] A. Tversky. Features of similarity. *Psychological Review*, 84: 327-352, 1977.
[65] E.M. Voorhees. The philosophy of information retrieval evaluation. *In Workshop of the cross-language evaluation forum for european languages*, 2001.
[66] H. Wang, et al. High expression levels of pyrimidine metabolic rate-limiting enzymes are adverse prognostic factors in lung adenocarcinoma: a study based on The Cancer Genome Atlas and Gene Expression Omnibus datastes. *Purinergic Signalling*, 16(3): 347-366, 2020.
[67] L.L. Wang, et al. Ontology alignment in the biomedical domain using entity definitions and context. *In BioNLP*, 2018.
[68] Z. Wang, A. Lachmann, A. Ma'ayan. Mining data and metadata from the gene expression omnibus. *Biophysical Reviews*, 11(1):103-110, 2019.
[69] Z. Wang, et al. Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nature Communications*, 7(1): 1-11, 2016.
[70] Z. Wu, M. Palmer. Verbs semantics and lexical selection. *In ACL*, 1994.
[71] D. Yin, et al. Ranking relevance in yahoo search. *In SIGKDD*, 2016.
[72] T. Zhang, et al. KIAA0101 is a novel transcriptional target of FoxM1 and is involved in the regulation of hepatocellular carcinoma microvascular invasion by regulating epithelial-mesenchymal transition. *Journal of Cancer*, 10(15): 3501, 2019.
[73] Y. Zhu, et al. GEOmetadb: powerful alternative search engine for the gene expression omnibus. *Bioinformatics*, 24(23): 2798-2800, 2008.