Toward Semantic Search for the Biogeochemical Literature

Joshua D. Eisenberg[1], Deya Banisakher[1], Maria Presa[1],
Kalli Unthank[2], Mark A. Finlayson[1], Rene Price[2], and Shu-Ching Chen[1]
[1] School of Computing and Information Sciences
[2] Department of Earth and Environment & Southeast Environmental Research Program
CREST Center for Aquatic Chemistry and the Environment
Florida International University
Miami, FL 33199
{jeise003, dbani001, mpres029, kunthank,
markaf, pricer, chens}@fiu.edu

Abstract

Literature search is a vital step of every research project. Semantic literature search is an approach to article retrieval and ranking using concepts rather than keywords, in an attempt to address the well-known deficiencies of keyword-based search, namely, (1) retrieval of an overwhelming number of results, (2) rankings that do not precisely reflect true relevance, and (3) the omission of relevant results because they do not contain the idiosyncratic keywords of the query. The difficulty of semantic search, however, is that it requires significant knowledge engineering, often in the form of conceptual ontologies tailored to a particular scientific domain. It also requires non-trivial tuning, in the form of domain-specific term and concepts weights. Here we present preliminary, work-in-progress results in the development of a semantic search system for the biogeochemical scientific literature. We report the following initial steps: first, one of the co-authors—a biogeochemistry expert—wrote a sample search query, and ranked the five most relevant articles that were returned for that query from a popular keyword-based search engine. We then hand annotated the five articles and the query with the Environmental Ontology (ENVO), an existing ontology for the domain. Critically, this pilot annotation revealed a number of missing concepts that we will add in future work. We then showed that a straightforward ontology distance metric between concepts in the search query and the five articles was sufficient to produce the expected ranking. We discuss the implications of these results, and outline next steps required produce a full-fledged semantic search system for the biogeochemistry scientific literature.

Keywords: Natural Language Processing, Semantic Search, Ontologies

1. Introduction

We all have had the experience of searching the scientific literature using a keyword-based search engine. You probably started with a general query, which returned thousands of articles that only tangentially related to your interests. Because no researcher would have time to even skim all the results, you returned to the original query, rewording it multiple times in different ways until highly relevant articles were ranked at the top of the search. These are long-known deficiencies of keyword-based search, namely: (1) retrieval of an overwhelming number of results, (2) rankings that do not precisely reflect true relevance, and (3) the omission of relevant results because they do not contain the idiosyncratic keywords of the query [26].

A long-proposed solution to this problem is *semantic search*, which uses concepts in the query rather than just keywords to drive document retrieval and ranking. Semantic search often leverages domain-specific knowledge, usually encoded in ontologies, to help rank the relevance of documents relative to a search query. Semantic search is difficult, however, because the required knowledge entails significant knowledge engineering or sophisticated natural language processing (NLP). Despite these problems, search engines today do boast high performance compared to prior decades precisely because they include minor semantic knowledge in their search algorithms; one approach, for example, is Latent Semantic Indexing (LSI), which uses synonyms and relationships between page headers, document titles, and content to assist ranking [16]. Nevertheless, we are still far from the full realization of true semantic search that uses deep semantic techniques fully integrated into back-end algorithms. For these reasons semantic search research is experiencing a rise in interest among various groups [4, 14, 16].

Here we present the first steps of work in progress aimed at implementing semantic search in a specialized domain—biogeochemistry—by reusing and integrating prior work on ontologies, semantic search, and NLP. The goal of this preliminary study is to show that, in principle, it is feasible to use an existing ontology, the Environment Ontology (ENVO) [6], to accurately encode domain-relevant concepts present in search queries and scientific articles and then use those concepts to correctly rank articles relative to the search query. We performed a pilot annotation of a test query and its five most relevant articles (as identified by our co-author, a biogeochemistry expert) which achieved high inter-rater reliability. We then devised a simple method for scoring articles relative to the query, based on the annotated concepts and their graph distances in ENVO, and showed that this method can reproduce the correct ranking, in contrast to keyword-based search engines as well as other baselines. The remainder of this paper is organized as follows: We first review related work on semantic search (§2). We next describe ENVO and the proposed complete system we seek to build, along with what we actually implemented for this work-in-progress report (§3). We then outline and discuss the pilot annotation study, the proposed ranking algorithm, and the results (§4). Finally, we discuss future directions (§5) and specify this work's contributions (§6).

2. Related Work

Concentrated work on semantic search started around 2000, when Heflin *et al.* built the Simple HTML Ontology Extensions (SHOE) search engine [12]. This search engine relied on manual tagging, via a markup language, of web pages with categories, relationships, and attributes drawn from an ontology. At the time of SHOE's release, NLP tools were unable to reliably extract ontology concepts from text and therefore the tagging process for documents was performed manually. Users queried the search engine by selecting concepts or other ontology features from a drop down menu. SHOE was the first example of the dominant idea of the 2000s, namely, that ontologies should be used to mark up web pages and create a *semantic web* to improve the performance of search algorithms [20].

This approach to semantic search sets the stage for our work, as we seek to build a system equivalent to a *fully automated* SHOE search engine for biogeochemical literature. In particular, because NLP methods and tools have advanced significantly since the release of SHOE, we also aim to go beyond earlier approaches by automatically extracting concepts and the underlying relationships from articles and queries. We will work on automating the extraction of ontological features from natural text found in articles from the biogeochemical domain.

Others working on semantic search in different domains have also leveraged advances in NLP in recent years. In 2012, Thomas *et al.* released GeneView, a semantic search system over PubMed articles [25], which provides entity specific search over 270,000 articles. GeneView uses specialized named entity recognizers (NERs) for different types of entities including genes, chemicals, and generic drug names. GeneView automatically tags each article or document with the entities it contains, resulting in the extraction of 194 million entities from PubMed.

Similarly, in 2015 the GATE Group at Sheffield University released *Mimir: Multiparadigm Indexing and Retrieval*, a semantic search system that scales to large data sets. The Mimir framework indexes documents using knowledge expressed in an ontology chosen by the user, and uses NLP tools that perform NER, entity linking, and semantic annotation. The system allows search over document structure, text, linguistic annotation, and semantic or ontological knowledge. As of 2015, Mimir was the only open source semantic search framework available, and we plan to leverage this project in our proposed system.

Closer to the biogeochemical domain, Hu *et al.* built a semantic search system for geospatial data in 2015 [13] which also uses automatic concept extraction: NERs trained on DBpedia [3] are used to tag geographic concepts and entities. Interestingly, that system also uses Labeled Latent Dirichlet Allocation (LLDA) to rank the metadata of different resources and harmonize different metadata formats. The ranking algorithm uses vector similarity for the query and the search domain of articles.

Finally, a key inspiration for our semantic search approach is BabelNet [18]. BabelNet includes a semantic search system that uses structured knowledge from Wikipedia and WordNet [8], in addition to the information gleaned from standard keyword search. Most novel to BabelNet's approach is that they created their own ontology, where WordNet senses and Wikipedia pages are used as ontology concepts. BabelNet starts by creating two graphs, one for Wikipedia and one for WordNet: hyperlinks between Wikipedia pages are used to instantiate links between concepts representing those pages, while semantic and lexical pointers are used to link WordNet senses. They then merge the two graphs to form one final ontology with over five million concepts and over 50 million links.

Stimulated by new semantic search engines entering the market, well-known search companies have also recently devoted attention to incorporating semantics into their search and ranking algorithms. Google, for example, introduced their Hummingbird search algorithm (or "conversational searches") in 2013, which emphasizes semantics, query context, document content, and page content more than earlier approaches [26]. Yahoo and Microsoft's Bing have employed similar measures to improve precision in their systems [24, 29].

3. System Design

Our proposed semantic search system for the biogeochemical literature will have several components, not all of which have been implemented for this paper. As mentioned previously, here we focus on (1) demonstrating the feasibility of encoding concepts in search queries and scientific articles using ENVO, our chosen domain-specific ontology, and (2) testing that, in principle, ENVO concepts can be used to correctly rank articles relative to a query. In this section we describe the ENVO ontology in detail, as well as outline the overall structure of the proposed system, indicating which components we have implemented for this paper.

3.1. The Environment Ontology

After a review of available domain-relevant ontologies, we determined that the most useful one for our purposes was the Environment Ontology (ENVO), a community-led, open ontology for various life science disciplines [6]. According to its creators, ENVO is an attempt at establishing a standard annotation scheme for several co-dependent or related disciplines, including, but not limited to, ecology, hydrology, environmental biology, and the geospatial sciences. ENVO contains concepts corresponding to a wide range of natural environments and environmental conditions. It is encoded in the Open Biomedical Ontologies (OBO) syntax, which is a subset of the Web Ontology Language (OWL). ENVO can be populated, managed, and maintained using the OBO-Edit ontology development tool.

ENVO, like many ontologies, is hierarchical in design. Three of its top-level, most developed branches are *environmental system*, *environmental feature*, and *environmental material*. It's hierarchical structure allows for it to include not only entities, but also higher-level relationships between various concepts, including many standard ontological relationships such as `is-a`, `part-of`, `contained-in`, `connects`, and `has-condition`. ENVO also contains scientific and domain-specific relationships such as `derives-from`, `input-of`, `output-of`, `has-habitat`, and `biomechanically-related-to`. Furthermore, the ontology boasts a well-connected graph of synonymy relationships, encoded using different granularities including `broad`, `exact`, and `narrow`.

ENVO has seen quite a bit of success in adoption and use. It has served as the foundation for the creation and expansion of a number of other ontologies, as well as applied in several annotation projects such as the International Census of Marine Microbes (ICOMM) and the International Nucleotide Sequence Database Collaboration (INSDC) [9]. Additionally, ENVO has been used in data retrieval and query-based systems such as the Genomic Metadata for Infectious Agents Database (GEMINA) [23], while the National Institute for Allergy and Infectious Diseases Bioinformatics Resource Centers (NIAID BRCs) employ ENVO in metadata formulation and manipulation [19].

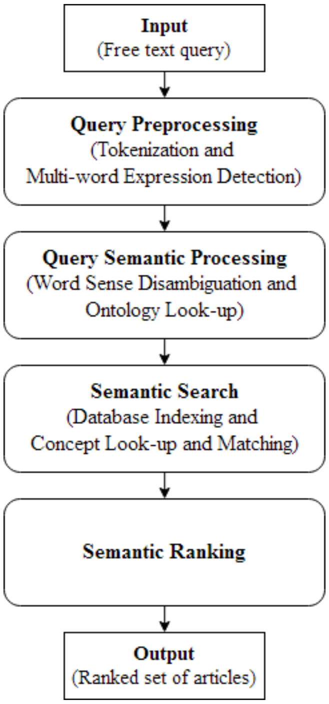

Figure 1. Architecture of the proposed semantic search system.

3.2. Overview of the Proposed System

In the work described here, we test the feasibility of two of the most important steps of the proposed semantic search system; however, there are four components in the full final system, as shown in Figure 1, and described below.

The first component will be query preprocessing. This component tokenizes the query and detects multiword expressions [15]. For example, the term *water* in isolation has one meaning, which is different from its meaning when embedded in the phrase *brackish water*, which is a multiword expression.

The second component will extract concepts from the query. In the work described here we had human annotators extract the concepts manually. In the proposed system, we will build NLP-based concept detectors that associate spans of texts with ontology concepts. This is a similar task to word sense disambiguation [1], except the system will decide between ontology concepts, not word senses.

The third component will perform the actual semantic search. The concepts found in the query are used to find relevant articles from a database of biogeochemical research articles. In the work described here our human annotators manually extracted the concepts from the five sample research articles under study (Table 1). In the proposed system, the ontology concept detectors will be used to index articles for the concepts they contain.

The fourth and final component will rank the articles retrieved during semantic search. The ranking algorithm used in this preliminary work is discussed in §4.2, and will serve as a starting point for developing a ranking algorithm robust enough for a larger corpus of research articles. While we anticipate that the final ranking algorithm will be more sophisticated than the one described here, we do expect that they will still share key similarities.

4. Feasibility Studies

The preliminary work described here consists of two parts: a pilot annotation study and tests of several possible ranking algorithms. For the pilot annotation, we manually applied concepts from the ENVO ontology to spans of text in a test query and the five most relevant scientific articles for that query (the test set). To identify a viable approach to ranking, we built a set of ranking algorithms that used the annotated concepts to produce a ranking on the test set. These two steps are discussed in turn below.

4.1. Annotation Study

The purpose of manually annotating concepts from the ontology is twofold: first, to show that the ontological concepts appear in the target texts, and, second, to show that it is possible to automatically rank articles when the concepts that appear in them are known. As noted, because developing NLP-based concept detectors is a non-trivial task, we wanted to test the utility of the ontology beforehand, as well as verifying that it is feasible to obtain a correct ranking for a small set of articles using those concepts.

One member of the team (KU) wrote a search query of interest to a current research problem in her lab:

> *Methyl-Mercury concentrations in Everglades water and sediment*

We ran this query through Google Scholar and she then identified, retrieved, and ranked the five most relevant articles from the search, extracted from the several hundred results returned. Importantly, several of the articles were not ranked near the top of Google's results, and were rather found many pages deep.

> *More than 20 years ago, Andren & Harriss (1973) measured relatively high % MeHg (MeHg as a percent of total Hg) in Everglades sediments, noting that samples from the Everglades were comparable to Hg-contaminated Mobile Bay sediments. [11, p. 328]*

Text Span	Concept	ID
Everglades sediments	sediment	2007
Everglades	peat swamp	189
Mobile Bay sediments	sediment	2007

Figure 2. Example sentence from article [11], page 328. Underlined portions of the text indicate spans that were associated with an ENVO concept; the table shows the associated ENVO concept ID.

The first four authors then annotated the query and the articles for concepts from ENVO. The query contained the following concepts: (1) *peat swamp*, (2) *sediment*, and (3) *water*. For each article, annotations were collected at the sentence level: we separated each article into a list of sentences where spans of text were tagged with ENVO concepts where they appeared.

To help us decide what concepts should be chosen for each span of text, we used Protégé [17] to search and explore ENVO. We recorded our annotations in a spreadsheet, where each row represented a sentence of text from the article, followed by columns representing the span of text containing the ENVO concept, the ID of the identified concept, as well as additional columns to capture text which seemed to represent concepts not present in the ontology. The most common missing concept we discovered during this process was *contaminant*. Figure 2 gives an example sentence from one of the test articles, along with the text spans which were associated with an ENVO concept.

The process of annotation was non-trivial, and involved several rounds of training, annotating, and revision of the annotation guidelines. Even for a relatively simple sentence as shown in Figure 2, numerous annotation decisions were needed. Below, we walk through this process phrase by phrase:

More than 20 years—This phrase does not need to be annotated, as it is a temporal expression referring to time period of the events mentioned later in the sentence.

. . . *Andren & Harris (1973)*—This phrase also does not need to be annotated, because it is a reference to a relevant article, and referring to the scientific literature isn't a concept in ENVO.

. . . *measured relatively high %*—This does not need to be annotated, as ENVO does not contain concepts related to specific chemical concentration levels.

. . . *MeHg*—This is the chemical formula for *methylmercury*, an environmental contaminant. The concepts of *contaminant* and *contamination* are not in ENVO. However, because this concept is relevant to the domain of interest, we did record these text spans and their related ideas so as to begin to build a set of concepts to expand ENVO in future work.

. . . *(MeHg as a percent of total Hg)*—Again, we identified the spans *MeHg* and *Hg* as the missing concept *contaminant*.

. . . *in Everglades sediments*—This phrase is tricky, because *Everglades* and *sediment* appear as individual concepts in ENVO, but when they appear in succession they form a multiword expression. *Everglades sediment* does not appear directly in ENVO. However, as it is presumably a subclass (or multiple subclasses) of sediments generally, we queried ENVO for the entity *sediment* (ENVO ID 2007), and examined its children for potential matches. *Sediment* has multiple children, namely, specific subtypes such as *lake sediment* or *contaminated sediment*. However, because there is no concept corresponding to the specific collection of different types of sediments that comprise the Everglades, we tagged this with the more general entity *sediment*.

. . . *noting that samples from the Everglades*—For this span, we first looked through ENVO to find a concept for *Everglades*. The closest concept is *peat swamp* (ENVO ID 189), which has no children, and so we tag this span using this concept.

. . . *were comparable to Hg-contaminated Mobile Bay sediments.*—For this span, we again tagged *Hg* as the missing concept *contaminant*. In the same way as above for *Everglades sediment*, the phrase *Mobile bay sediments* was tagged with the general concept *sediment*.

The first four co-authors served as the annotators for this pilot annotation. We annotated the first 50 sentences of the first article [27] cooperatively to develop the annotation guidelines, while each annotator annotated the remaining 130 sentences individually so as to allow us to calculate inter-rater reliability. This produced a Cohen's κ of 0.57, which is "moderate to substantial" agreement [2]. We then assigned each of the annotators one of the four remaining articles for annotation [11, 5, 10, 7].

In Table 1 we present statistics on the test set. The articles have an average of 4,114 tokens, 165 sentences, 21 unique ENVO concepts. In §4.2 we discuss how our ranking algorithm accounts for this variance, and we detail how the scores are calculated.

4.2. Proposed Ranking Algorithm

The result of annotation was a list of concepts for each article and the position of the concepts within the article. This allowed us to calculate the number of times a concept appears. The ranking algorithms we tested used this concept list in comparison with the query concept list to produce a ranking.

We were able to design an algorithm that correctly ranks all five articles (in Table 2 we call this algorithm *Graph Search*). The psuedocode for this algorithm is presented in Algorithm 1.

Algorithm 1 Pseudocode for Graph Search Algorithm

```
for paper : database do
    $score \leftarrow 0$
    for qc : query.concepts do
        for pc : paper.concepts do
            $dist \leftarrow distance(qc, pc)$
            $dist \leftarrow dist \cdot tfidf(pc)$
            $score \leftarrow score + dist$
        end for
    end for
    $paper.score \leftarrow score \cdot (1 + \alpha \cdot D)$
end for
```

The idea behind this algorithm is the most highly ranked articles should contain concepts that are closely related to concepts in the query. The algorithm works as follows. First, we model semantic relatedness by determining whether concepts are connected by an ancestor-descendant chain in the ontology. We count the number of links in this chain, and then weight this count by the tf-idf of the concepts, as computed over the set of all articles in the database (which here is the same as the test set). We then summed these scaled counts to produce a similarity score for each article.

Second, because article length will correlate with the number of concepts (all other things being equal), the similarity measure as described so far will also correlate with length. We need to correct for this effect: just because an article is short doesn't mean it shouldn't be highly ranked. Moreover, it is also important to consider how much of the article is actually relevant to the domain. Research articles can be about multiple topics, and some of these topics may not be relevant to biogeochemistry, or covered by ENVO. We propose that both of these problems can be addressed by scaling the similarity measure by a factor dependent on the concept density in relevant sections. In this approach, articles with higher densities concepts should have their similarity score boosted, and articles with lower densities should have their score suppressed. We implemented this by computing the ratio of total concept mentions in an article to

Rank	Title	Citation	Tokens	Sentences	Relevant Sents.	Unique Concepts	Raw Sim. Score	Scaled Sim. Score
1	Mercury in the Aquatic Environment . . .	[27]	5,081	162	162	26	1.39	1.19
2	Methylmercury Concentrations . . .	[11]	4,295	183	183	26	0.97	0.84
3	Sulfide Controls on Mercury Speciation . . .	[5]	4,133	168	114	13	0.43	0.73
4	Sulfate Stimulation of Mercury Methylation . . .	[10]	3,642	160	160	18	0.75	0.51
5	Effect of Salinity on Mercury Activity . . .	[7]	3,421	150	150	22	0.50	0.45
	Average		4,114	165		21		

Table 1. Articles in the test set. Listed are the number of tokens in each article, the number of sentences overall, and the number of sentences in conceptually relevant sections (note, only article 3 had irrelevant sections, resulting in 114 sentences in relevant sections). The last two columns are described in §4.2.

the number of sentences in *relevant* sections. We say that a section is *relevant* if it contains at least one ENVO concept. We computed the mean and standard deviation of the densities across the test set, and then computed the number of standard deviations from the mean for each article (D). We then multiplied the raw similarity score described above (second-to-last column of Table 1) by the scaling factor $1 + \alpha D$, where in this case the tuning parameter α was set to 2.5. This value is likely fairly specific to this test set, and our future work will explore the correct range of values for this parameter, and their effects on the search. The resulting scaled similarity score is shown in the last column of Table 1.

The complexity of the algorithm is dependent on the number of papers in the database (p), the number of concepts in the search query (q), and the number of concepts in each paper (c). The computation relies on looks up of the distance between ontology entities, and the tf-df of the paper concepts, both of which can be precomputed. Thus the time complexity of the algorithm is $O(p \cdot q \cdot c)$.

4.3. Preliminary Results and Discussion

We compared the Graph Search on the full document text approach to five other ontology-based methods. First, we compared with simple tf-idf weighting applied directly to concepts appearing in the full text (Concept tf-idf). We used the traditional tf-idf algorithm [22] on the list of unique concepts in each document. The tf-idf weights for each concept were summed for each article, and these sums were used to rank the articles. Second, we compared to a simple concept counting algorithm, where we count the number of times concepts in the search query appear in an article (Concept Counting). We also compared with the Graph Search, Concept tf-idf, and Concept Counting approaches applied just to the article abstracts.

Table 2 shows a comparison of the ranking results for these different ranking algorithms. The articles in the table are listed in their correct order provided by our domain expert (KU).

Additionally, we collected rankings from four state-of-the-art or easily available scientific literature engines: (1) Google Scholar, (2) Microsoft Academic, (3) Web of Science, and (4) our university library's in-house article search engine. The in-house search engine searches over one thousand databases such as IEEE, PubMed, and Elsevier using standard keyword-based search. To obtain rankings from these search engines we provided the test query to each and looked for each of the identified articles in the results.

The graph search algorithm generated the correct ranking for each article. The Concept tf-idf and Concept Counting approaches were less successful; the former, however, did rank four of the five articles correctly. Although they use semantic knowledge these algorithms are similar to classical keyword-based search, since they are making decisions based only on the presence or absence of concepts that occur in each article.

For the abstract only tests, the graph search algorithm was able to rank the top two articles correctly, while misplacing the last three. For the two of the articles the Concept Counting algorithm ranked them the same

Furthermore, the graph search algorithm performed better than the available search engines for ranking this test set. Google Scholar was able to retrieve all five articles and rank three of them correctly (relative to other articles). However, Microsoft Academic did not retrieve four articles, and our university library's system did not retrieve any of the articles, even though all of those articles were present in their

	Graph Search	Concept tf-idf	Concept Counting	Graph Search	Concept tf-idf	Concept Counting	Google Scholar	Microsoft Academic	Web of Science	University Library
Article 1	1	2	4	3	5	5	3	-	3	-
Article 2	2	3	5	1	3	1,2	1	-	1	-
Article 3	3	4	3	4	1	1,2	2	1	2	-
Article 4	4	5	1	2	2	3	4	-	5	-
Article 5	5	1	2	5	4	4	5	-	4	-
	Full Article			**Abstract Only**			**Baseline**			

Table 2. Ranking results for baseline and models employed in full articles and abstract only sets

databases. Web of Science was able to retrieve all the articles, but with an incorrect ranking.

Finally, we note that there is nothing especially specific to the biogeochemical domain in our approach. Thus we expect that the Graph Search algorithm could be applied to other domains, as long as one has relevant ontologies and methods to automatically extract ontology concepts from text.

5. Future Directions

We have demonstrated the feasibility of several important steps in the development of semantic search for the biogeochemical literature. In future work, we plan to build an end-to-end system that implements full, automatic semantic search for the domain. For this pilot study, several parts of the system were simulated by human computation, such as the extraction of concepts and relevant sections. Next steps will include collating an extensive database of biogeochemistry articles, refining and extensively testing our semantic similarity measure, creating automatic concept extractors for text, and expanding the ENVO ontology with missing concepts. We discuss each of these in turn.

5.1. Database

A key step in building the proposed system is to develop a database of biogeochemistry scientific articles. For this we propose to begin with data provided by Elsevier, which in 2014 released a new Application Programming Interface (API) to make it easier to text-mine scholarly articles [28]. The articles are provided in XML format, where information such as title, authors, and content are explicitly tagged making the document easy to parse, following Elsevier's principle for most of its content which is "XML first." We will extract from Elsevier's databases the articles related to biogeochemistry.

5.2. Evaluating and Enhancing the Ranking Algorithm

As previously discussed, there are many different approaches to semantic search. In this paper, we presented a simple method for ranking tailored to small test set. However, it is clear that we will need to engage in much more significant evaluation of our ranking algorithm to demonstrate that it works for the significantly larger task of ranking thousands of articles across many different queries. No doubt, this work evaluation will lead to many refinements and additions. For instance, in this study we only considered ancestors-descendant relationships between concepts. In our future work we will consider more complex paths between concepts, as well as other attributes that are encoded in the ontology. We will also need to expand the database from which we calculate our tf-idf weights. Concepts that appear frequently in the literature, like *water* and *sediment*, can overwhelm the ranking algorithm because of their frequent use.

5.3. Developing Automatic Concept Extractors

We also plan to automate the task of extracting ENVO concepts from text, using state-of-the-art NLP approaches. To assist in this we plan to expand our set of annotations with ten more biogeochemical research articles (approximately 40,000 words). The ten additional articles will each be independently annotated by two annotators, and the conflicts in the annotations will be resolved by adjudication. This will result in a larger corpus of gold-standard ENVO annotations, and allow us to measure more accurate agreement statistics on this task. These data will allow us to evaluate our to-be-developed entity and concept extractors, and also help in improving our ranking algorithm.

5.4. Ontology Expansion

Although the ENVO ontology contains a large number of concepts and entities related to the environment, the ontology by design leans towards biology rather than hydrology. For example, the ontology does not contain concepts for *contaminant*, *methylation*, and *trophic*, all of which are extremely common in the broader biogeochemical domain. For this reason, we plan to extend ENVO by adding concepts and leveraging existing mappings between ENVO and other related ontologies, e.g., to the Semantic Web for Earth and Environmental Terminology (SWEET) ontologies [21].

6. Contributions

We have described preliminary, work-in-progress results exploring the feasibility of using the ENVO ontology to enable semantic search for the scientific literature in the biogeochemical domain. We confirmed that ENVO does capture many important concepts expressed in these articles, and demonstrated a straightforward ranking algorithm that correctly ranks the articles in our test set relative to the test query. We plan to continue this work by constructing a database of biogeochemical articles, building and evaluating a more sophisticated ranking algorithm, creating automatic concept extractors for ENVO tested using more annotated data, and expanding the ENVO ontology with important missing concepts.

7. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. HRD-1547798, which was awarded to Florida Internation University as part of the Centers of Research Excellence in Science and Technology (CREST) Program. This is contribution number 838 from the Southeast Environmental Research Center in the Institute of Water and Environment at Florida International University.

References

[1] E. Agirre and P. Edmonds. *Word Sense Disambiguation*. Springer, Dordrecht, The Netherlands, 2007.

[2] R. Artstein and M. Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web*, pages 722–735, 2007.

[4] H. Bast, B. Buchhold, and E. Haussmann. Semantic Search on Text and Knowledge Bases. *Foundations and Trends in Information Retrieval*, 10(2-3):119–271, 2016.

[5] J. M. Benoit, C. C. Gilmour, R. P. Mason, and A. Heyes. Sulfide Controls on Mercury Speciation and Bioavailability to Methylating Bacteria in Sediment Pore Waters. *Environmental Science & Technology*, 33(6):951–957, 1999.

[6] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, and S. E. Lewis. The environment ontology: contextualizing biological and biomedical entities. *Journal of Biomedical Semantics*, 4(1):43, 2013.

[7] G. C. Compeau and R. Bartha. Effect of Salinity on Mercury-Methylating Activity of Sulfate-Reducing Bacteria in Estuarine Sediments. *Applied and Environmental Microbiology*, 53(2):261–265, 1987.

[8] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 2000.

[9] D. Field, L. Amaral-Zettler, G. Cochrane, J. R. Cole, P. Dawyndt, G. M. Garrity, J. Gilbert, F. O. Glöckner, L. Hirschman, and I. a. Karsch-Mizrachi. The Genomic Standards Consortium. *PLoS Biology*, 9(6):e1001088, 2011.

[10] C. C. Gilmour, E. A. Henry, and R. Mitchell. Sulfate Stimulation of Mercury Methylation in Freshwater Sediments. *Environmental Science & Technology*, 26(11):2281–2287, 1992.

[11] C. C. Gilmour, G. Riedel, M. Ederington, J. Bell, G. Gill, and M. Stordal. Methylmercury concentrations and production rates across a trophic gradient in the northern Everglades. *Biogeochemistry*, 40(2-3):327–345, 1998.

[12] J. Heflin and J. Hendler. Searching the Web with SHOE. In *Proceedings of the AAAI-2000 Workshop on AI for Web Search*, pages 35–40, 2000.

[13] Y. Hu, K. Janowicz, S. Prasad, and S. Gao. Metadata Topic Harmonization and Semantic Search for Linked-Data-Driven Geoportals: A Case Study Using ArcGIS Online. *Transactions in GIS*, 19(3):398–416, 2015.

[14] V. Jindal, S. Bawa, and S. Batra. A review of ranking approaches for semantic search on web. *Information Processing & Management*, 50(2):416–425, 2014.

[15] N. Kulkarni and M. A. Finlayson. jMWE: A Java Toolkit for Detecting Multi-Word Expressions. In *Proceedings of the 8th Workshop on Multiword Expressions*, pages 122–124, Portland, OR, 2011.

[16] W. Li, M. F. Goodchild, and R. Raskin. Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth*, 7(1):17–37, 2014.

[17] M. A. Musen. The Protégé Project: A Look Back and a Look Forward. *AI matters*, 1(4):4–12, 2015.

[18] R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[19] The National Institute for Allergy and Infectious Diseases (NIAID), Microbiology and Infectious Diseases Resources, DMID Metadata Standards Core Sample. `https://www.niaid.nih.gov/research/dmid-metadata-standards-core-sample`, 2017. Retrieved on May 9, 2017.

[20] N. F. Noy, M. Sintek, S. Decker, M. Crubézy, R. W. Fergerson, and M. A. Musen. Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16(2):60–71, 2001.

[21] R. Raskin. Semantic Web for Earth and Environmental Terminology (SWEET). Technical report, Jet Propulsion Laboratory, 2003.

[22] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. 1986.

[23] L. M. Schriml, C. Arze, S. Nadendla, A. Ganapathy, V. Felix, A. Mahurkar, K. Phillippy, A. Gussman, S. Angiuoli, E. Ghedin, O. White, and N. Hall. GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database. *Nucleic Acids Research*, 38(suppl 1):D754–D764, 2010.

[24] T. Seymour, D. Frantsvog, and S. Kumar. History of Search Engines. *International Journal of Management and Information Systems*, 15(4):47, 2011.

[25] P. Thomas, J. Starlinger, A. Vowinkel, S. Arzt, and U. Leser. GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Research*, 40(W1):W585–W591, 2012.

[26] D. Tümer, M. A. Shah, and Y. Bitirim. An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, MSN and Hakia. In *Internet Monitoring and Protection, 2009. ICIMP'09. Fourth International Conference on*, pages 51–55. IEEE, 2009.

[27] S. M. Ullrich, T. W. Tanton, and S. A. Abdrashitova. Mercury in the Aquatic Environment: A Review of Factors Affecting Methylation. *Critical Reviews in Environmental Science and Technology*, 31(3):241–293, 2001.

[28] R. Van Noorden. Elsevier opens its papers to text-mining. *Nature*, 506(7486):17–17, 2014.

[29] D. Yin, Y. Hu, J. Tang, T. Daly, M. Zhou, H. Ouyang, J. Chen, C. Kang, H. Deng, C. Nobata, J.-M. Langlois, and Y. Chang. Ranking Relevance in Yahoo Search. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 323–332. ACM, 2016.