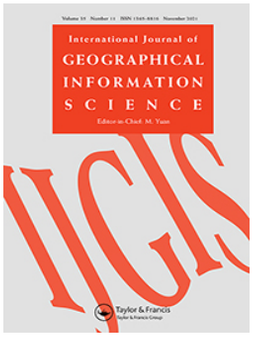

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

GIS-KG: building a large-scale hierarchical knowledge graph for geographic information science

Jiaxin Du, Shaohua Wang, Xinyue Ye, Diana S. Sinton & Karen Kemp

To cite this article: Jiaxin Du, Shaohua Wang, Xinyue Ye, Diana S. Sinton & Karen Kemp (2021): GIS-KG: building a large-scale hierarchical knowledge graph for geographic information science, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2021.2005795

To link to this article: https://doi.org/10.1080/13658816.2021.2005795

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2021.2005795

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

GIS-KG: building a large-scale hierarchical knowledge graph for geographic information science

Jiaxin Du[a], Shaohua Wang[b], Xinyue Ye[a], Diana S. Sinton[c,d] and Karen Kemp[e]

[a]Department of Landscape Architecture and Urban Planning, Texas A&M University, College Station, X, USA; [b]Ying Wu College of Computing, New Jersey Institute of Technology, Newark, NJ, USA; [c]University Consortium for Geographic Information Science, USA; [d]College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, USA; [e]Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angles, CA, USA

ABSTRACT

An organized knowledge base can facilitate the exploration of existing knowledge and the detection of emerging topics in a domain. Knowledge about and around Geographic Information Science and its associated system technologies (GIS) is complex, extensive and emerging rapidly. Taking the challenge, we built a GIS knowledge graph (GIS-KG) by (1) merging existing GIS bodies of knowledge to create a hierarchical ontology and then (2) applying deep-learning methods to map GIS publications to the ontology. We conducted several experiments on information retrieval to evaluate the novelty and effectiveness of the GIS-KG. Results showed the robust support of GIS-KG for knowledge search of existing GIS topics and potential to explore emerging research themes.

ARTICLE HISTORY
Received 24 September 2020
Accepted 9 November 2021

KEYWORDS
Geographic information science (GIS); ontology; knowledge graph; information retrieval; natural language processing

1. Introduction

Geographic Information Science and its associated system technologies (GIS) comprise an extensive field of knowledge, making it challenging to organize the content in a comprehensive digital manner. Diverse research from this cross-domain discipline includes geovisualization, cartography, spatial cognition, and spatial analytical methods within the physical, environmental, and social sciences, and extends to spatial approaches within computer science. The inter- and multi-disciplinary nature of the research involves not only the fundamentals of spatial science but also its applications in other domains. Aspects of GIScience, its related mapping technologies, and its diverse applications are taught widely across higher education institutions, and the rapid development of the technologies and new application areas contribute to the challenges of keeping curricula current.

A 'body of knowledge' consists of concepts, terms, approaches, and activities associated with a discipline, most often defined by professional practitioners of the domain (Bourque and Fairley *et al*. 2014). Maintaining a discipline's body of knowledge in a form

CONTACT Xinyue Ye xinyue.ye@tamu.edu
Supplemental data for this article can be accessed here.

that facilitates open access and review allows its fundamental scientific questions and answers to be transparent and accessible. A clearly defined body of knowledge supports a research community as it explores the ways in which new knowledge evolves, and provides for practical applications in education and workforce development.

In the realm of GIS, the first formal compilation of a Body of Knowledge was an effort led by the University Consortium for Geographic Information Science (UCGIS) and published as a 160-page paperback book in 2006 by the Association of American Geographers (DiBiase *et al*. 2006). Defining and organizing the content began with multiple iterative knowledge-gathering and brainstorming sessions across a broad higher education community, though it was ultimately completed by the small set of individuals listed as editors. At that time, no systematic plans existed for how the content would be maintained, curated, or updated. In 2016, the content was shifted to a digital platform to facilitate access and revisions (gistbok.ucgis.org). Its continued development is now modeled after the online *Stanford Encyclopedia of Philosophy* (Zalta *et al*. 1995), in which an editorial board oversees the production of entries authored by individual scholars who are typically academic experts in the topic.

Given the highly specialized and rapidly evolving nature of GIS, students of the science and its technologies are challenged to acquire holistic exposure to the breadth of the discipline. Indeed, this domain cannot be comprehensively described or analyzed through traditional analog efforts, especially when knowledge collections exist only on platforms that lack exploratory visualization functions, with little capacity for regular updates and revisions. Building on the original UCGIS Body of Knowledge framework, other organizations have developed specialized or expanded versions. These include the US Geospatial Intelligence Foundation's (USGIF) Essential Body of Knowledge, compiled by and for the intelligence and military community (Wang *et al*. 2020), and the recently released European body of knowledge for earth observation and geographic information (EO4GEO). Each of these are described in more detail later.

Our research aim is to improve domain-specific information retrieval systems for the broad discipline of geographic information science and its technologies. Tools such as Google Scholar (Martín-Martín *et al*. 2018, Gusenbauer 2018), Semantic Scholar (Fricke 2018) and Microsoft Academics (Harzing and Alakangas 2017) scan the entire population of published research during their searches. While these tools function as powerful document retrieval services, they do not provide focused insights nor facilitate understanding about how new knowledge is incorporated into an existing framework. This motivated us to pursue a new approach to knowledge organization and dissemination that still leverages the breadth of existing collections. First, we merged several collections of GIS-focused knowledge into a single ontology to support more extensive knowledge building informed by artificial intelligence (AI) and knowledge graphs.

Knowledge graphs are both an emerging paradigm and a technology stack that allows re-visioning of how knowledge is represented by highlighting the connections between concepts and properties. These graphs combine AI technologies and semantics data to represent densely interconnected statements derived from heterogeneous sources across domains in a manner that is readable by humans and machines (Janowicz 2021). A knowledge-graph consists of two parts: entities and the relationships between those entities. Because knowledge graphs address connections among arbitrary entities and their properties, they support complex and seamless crosswalks that can be conceptually

defined and explicitly represent identity and equivalence relationships between individuals and classes. These schemata are called ontologies, and their rich axiomatization supports semantic interoperability and machine reasoning. The ability of knowledge graphs to handle diverse and even contradictory ontologies is one of their core strengths. Moreover, having our GIS ontology derived from multiple complementary sources will produce a knowledge graph optimized for further input of emerging topics across the dynamic, evolving extent of GIScience, GIS, spatial science, and their diverse applications.

Our GIS-focused knowledge graph differs from other ontology-based bodies of knowledge in its use of state-of-the-art natural language processing technologies to map the semantic similarities among and between concepts and research activities. This mitigates one major concern with knowledge graphs applied to existing systems: how they accommodate heterogeneous concepts within a common space. We also developed computational techniques that support contextualization of domain concepts and research publications. This was particularly important because we aimed to capture GIS knowledge as completely as possible, by integrating knowledge from multiple primary sources. The result will accelerate the exploration of knowledge about GIScience, GIS, and its current and future applications, to benefit the higher education community and its critical connections with practitioners, employers, and clients. By having academic research more readily accessible and appropriately linked with relevant work competencies and skills, it makes it easier for practitioners to search for knowledge in their domain.

The contributions of this paper are:

(1) **Expansion of GIS Knowledge**. Through a knowledge fusion approach, we defined a GIS knowledge graph (GIS-KG) that reveals heterogeneous relationships between GIS concepts and diverse source materials (including existing collections of GIS knowledge and competencies, as well as a vast number of scientific publications). The fusion itself was based both on semantic similarity and domain expert knowledge. This type of semi-automatic fusion is used for identifying related entities when merging multiple bodies of knowledge. The structure and semantic meanings of the merged ontology supports the organization of the input materials, resulting in a broader capture of knowledge across the fields of GIS.
(2) **Novel approaches**. We designed a novel deep learning-based approach to support the knowledge fusion framework. Our approach took advantage of advanced deep learning models to measure and understand the semantic similarities between the ontology and published research papers. We further extended this approach to create the novel GIS knowledge search system.
(3) **Extensive evaluation and benchmark data**. We evaluated the robustness of the GIS-KG by using specific information retrieval methods for GIS knowledge. By using the GIS-KG for information searches, we returned results that were almost 20 times more accurate and relevant than other internet-based searches. Such advances will become new benchmarks for retrieval of GIS knowledge.

Organization of this paper. Section 2 presents related work, Section 3 is the overall structure of the project, Section 4 introduces how we created the new GIS ontology by merging existing knowledge graphs, Section 5 introduces how we collected knowledge materials and the preliminary study on the data, Section 6 presents the ontology and

knowledge material fusion methodology for the final GIS-KG, Section 7 uses an information retrieval application to evaluate the knowledge graph we built in this research, Section 8 discusses the limitations and future direction of this work, and Section 9 is the conclusion.

2. Related work

There have been several efforts to systematically organize and study GIS knowledge, including bibliographic studies and the production of knowledge graphs from scratch by experts.

2.1. Bibliometric analysis in GIS

Conducting a bibliometric analysis involves studying a vast collection of research papers about a scientific topic in order to develop an overview of it. Without the aid of AI, such efforts are limited to young scientific fields with a limited number of published research papers. In 1991, Professor Duane Marble began to build a GIS Master Bibliography, aiming to improve the accessibility of GIS literature to the public (Marble 2000). Esri later became the curator of the bibliography and incorporated it into an online GIS Bibliography (Marble 2001). In the early 1990 s, Al-Taha *et al*. (1994) produced a bibliography on spatiotemporal databases and a collection of basic statistics for the data.

More recently, Biljecki (2016) analyzed the publications in GIS related journals from 2000 to 2014 based on output volume, citations, national output and efficiency. Siabato *et al*. (2014) studied publications relating to temporal GIS and built a topic-based visualization. Such research paper knowledge bases have been used in studying the intellectual domain of GIS (Skupin 2014) and its global impact (Zhan *et al*. 2014). GIS publications in applied domains have also been collected, including public health (Marble 2000), social justice (Cochrane *et al*. 2017), web mapping (Haklay *et al*. 2008), and human geography (Zhong *et al*. 2015). While these reviews are helpful, they are necessarily dated and limited in scope. Our collection aims for more comprehensive coverage of GIS, and in a manner that can be readily updated.

Systematically monitoring emerging topics as they appear in the GIS literature is also daunting to accomplish without the aid of AI. Research tends to focus on trends or new techniques in one area alone, such as spatial multicriteria analysis (Malczewski and Jankowski 2020), or susceptibility mapping (Ghorbanzadeh *et al*. 2020). Thus, it is a persistent challenge for GIS curricula in higher education to be comprehensive or current (Wikle and Sinton 2020). Only through AI-enhanced techniques can the already-large and expanding field of GIS and its applications be monitored. Even the Open Geospatial Consortium (OGC) has begun to use AI to inform its observations of technology trends (https://www.ogc.org/OGCtechExplorer).

2.2. GIS bodies of knowledge and ontological work

The 2006 UCGIS GIS&T Body of Knowledge (BoK) has served as the foundation or model for several other bodies of knowledge, including the Essential Body of Knowledge of the US Geospatial Intelligence Foundation (USGIF), compiled by and for the intelligence and

military community (Wang *et al*. 2020), and the recently released European body of knowledge for earth observation and geographic information (EO4GEO). Additional efforts at organizing GIS knowledge have focused on GIS workforce development. In the US, several groups have produced enumerations of competencies and tasks that GIS professionals (e.g. technicians, analysts, and managers) would be required to maintain or complete for their work. Two examples are the Department of Labor's Geospatial *Technology* Competency Model (DiBiase *et al*. 2010) as well as the Geospatial *Management* Competency Model (Babinski 2012, Johnson 2019). In both cases, their content was developed by soliciting and documenting the practices and activities of current professionals in the field.

Less frequently has there been an explicit reliance on or production of a GIS ontology. The 2006 GIS&T BoK was transformed by Ahearn *et al*. (2013) to a semantic network relying on an ontology to enable semantic referencing and other applications, but this has now been absorbed into proprietary, commercial research activities (bigknowledge.net). Tomaszewski and Holden (2012) used the same original BoK content to explore ontological connections between GIS and IT for curricular pursuits. The EO4GEO project started with the UCGIS BoK and the Ahearn *et al*. (2013) ontology to create a BoK that focuses on applied skill sets (Vandenbroucke and Vancauwenberghe 2016). The result has since been further expanded to include additional emerging topics in GIS and connect with the field of Earth Observation (EO), hence its name as the EO4GEO BoK (Hofer *et al*. 2020).

Most of these earlier efforts relied extensively – if not exclusively – on domain experts to designate content. Such an approach can be valid and produce robust results, which is why our GIS-KG merged these existing expert-knowledge collections as its first step. We then went further by having the content merged to produce a Knowledge Graph that supports targeted GIS information searches.

2.3. Information retrieval systems

Behind the curtain of information retrieval tools is a necessary process called embedding that maps words or phrases with vectors of real numbers so that the text can be recognized by a computer. The similarity between queries and documents can then be calculated using a cosine distance or more sophisticated measurements that speed the retrieval process. Different types of embedding processes affect the quality of information systems implementation. The traditional method for information retrieval from text documents is called the bag-of-words approach (Manning *et al*. 2008), in which only exact keyword matches are considered correct. The embedding-of-words method takes word co-appearance into consideration (Bojanowski *et al*. 2017) and it allows similar words (e.g. synonyms) to qualify as correct matches. The embedding-of-sentences method produces more contextual information about the text (Devlin *et al*. 2018) and can even represent semantic meanings. Embedding-of-sentences is usually calculated by pre-trained language models (Devlin *et al*. 2018, Dai and Callan 2019). For example, the BERT model uses the attention mechanism (Vaswani *et al*. 2017) to aggregate semantic information in sentences to a numerical vector. With a few modifications, embedding-of-sentence models have achieved the state-of-the-art results in semantic matching tasks

(Adhikari *et al*. 2019) and perform even better when trained on domain-specific texts (Wang *et al*. 2019). This motivated us to modify the embedding models and adapt the general models to the GIS domain to improve search performances.

2.4. Research with knowledge graphs

Research with and about knowledge graphs is becoming more common. The process of merging separate bodies of knowledge, defined as entity resolution or graph alignment, has been well documented (Trisedya *et al*. 2019). Graph embedding techniques have been developed to deal with the sparsity of information in knowledge graphs, by resolving vagueness and uncertainty at the conceptual level (Mai *et al*. 2020bb), and inferring new entities in the Knowledge Graph embedding space based on existing ones (Mai *et al*. 2020a). Those emphasized using knowledge graphs to retrieve spatial information, while our work focuses on the knowledge about geographic information science and its technologies more broadly.

By creating a robust and broad knowledge graph, effective and successful searches for additional information and recommendations for related resources can be supported. Knowledge search systems can be built directly on knowledge graphs to find related topics using 3D visualization (Li *et al*. 2019). Such systems benefit from massive pre-trained language models such as deep bidirectional transformers (Devlin *et al*. 2018) and auto-regressive language models (Brown *et al*. 2020) that take advantage of feature embedding. Besides text data, the language models can also improve semi-structured data search results (Herzig *et al*. 2020).

Knowledge graphs are heavily used in question and answering (QA) systems (Mai *et al*. 2020a, Ye *et al*. 2021). The question answering process can be seen as an effort to minimize the entropy of user intention, so that the QA system can better anticipate what the user wanted and reduce uncertainty in the answers (Wang *et al*. 2021). Questionnaire is a general approach to alleviate the cold start problem for information retrieval and recommendation systems (Fan *et al*. 2019). Other processes including collecting extra user information and asking follow-up questions serve as query expansion (MacAvaney *et al*. 2018). The extra information in the embedding helps lower the entropy of user intention, resulting in more reliable search outcomes.

3. Overview of this research

Forging a common comprehensive knowledge graph for a field as young and dynamic as GIS is a complicated task but the result is powerfully useful. Our goal was to create a large-scale hierarchical knowledge graph in which the core of research, education and professional activities within the GIS domain can be discovered and explored. This will serve as a timely and foundational scientific knowledge base for the field.

Building this knowledge graph required completion of the following steps shown in Figure 1:

- **Step 1**. Forming a unified ontology for GIS. The ontology built in this step is the backbone of the knowledge graph. We first identified existing bodies of knowledge and competency models and merged them into a single hierarchical structure. We

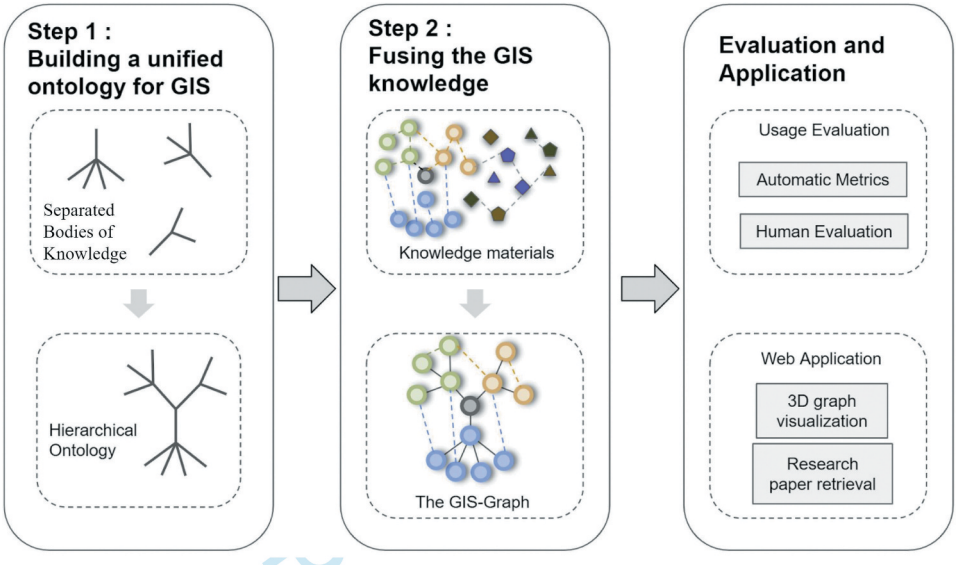

Figure 1. Overall structure of this paper. Step 1 is to merge different bodies of knowledge and build a unified ontology for GIS. Step 2 is to fuse the GIS knowledge materials with the ontology and form a comprehensive GIS-KG. Then, we can build applications and evaluate the GIS-KG.

utilized text similarity and structural information for automatic entity resolution and alignment. A manual check was performed after the merge to refine the ontology. Details of this step can be found in Section 4

- **Step 2**. Enriching the GIS knowledge. We then collected a vast number of GIS-related research papers and related information from open sources (Tang *et al*. 2008, Sinha *et al*. 2015). The ontology we had developed in the first step provided sufficient information for us to separate out and organize the GIS-focused research publications. Besides a paper's title and abstract, we also captured citation and venue features as additional information. We used parallel computing technologies to process this large amount of text-based, unstructured data, and ran basic analyses on the resulting data set. The result was a novel deep-learning-based approach that matched papers to the ontology and formed the final GIS-KG. Details of this step can be found in Sections 5 and 6.
- **Step 3. Evaluation and application**. To evaluate the utility of the GIS-KG, we built several applications to illustrate its benefits. We conducted a series of experiments on information retrieval tasks in which publications were returned based on user queries. The GIS-KG enhanced both traditional and advanced search methods. Finally, a web-based system was developed for users to further explore the GIS-KG, which is presented in Section 7.

4. Building a unified ontology for GIS

As mentioned earlier, several bodies of knowledge and other comprehensive collections of GIS-specific competencies exist in the field of geographic information science and technology, but each was created for different purposes by different organizations, and therefore has different elements within its content. Fortunately, the one type of informational knowledge that exists across these collections are learning objectives or stated competencies, so linkages could be ontologically constructed from these. While learning objectives and competencies are not always identical in format or intent, both are text statements and typically start with a verb whose activity can be quantitatively or qualitatively measured for assessment purposes. Each describes what a student or practitioner should be able to know or do. For example, one of the learning objectives for the UCGIS GIS&T BoK topic of 'Overlay' is 'Demonstrate why the geo registration of datasets is critical to the success of any map overlay operation'.

4.1. Sources for the GIS ontology

We used the following bodies of knowledge and competency models to build our hierarchical ontology (see Table 1):

- UCGIS GIS&T BoK (DiBiase *et al*. 2006) The aims of the 2006 BoK were to document the domain of Geographic Information Science and its associated technologies, but its original 2006 content consisted only of topic titles and learning objectives, grouped and hierarchically arranged within knowledge areas. The current GIS&T BoK continues to be developed and expanded so that each Topic consists of a much lengthier and detailed descriptive narrative in addition to learning objectives, but only about 40% of the Topics had been completed in the March 2020 version that we extracted to utilize in this study. Thus, our input included Topics that had been expanded as well as Topics from 2006 having only learning objectives. The context we extracted included 10 Knowledge Areas (first level), 96 Units (second level), 401 topics (third level) and 1467 learning objectives (fourth level). This is shown in Table 1.

Table 1. Key information about the bodies of knowledge that we used to create the new ontology.

Source	Caretaker	Content	Version
GIS&T body of knowledge (BoK)	University Consortium for GIScience	Knowledge Areas, Units, Topics, Learning Objectives, Narratives	online accessed in March 2020
Essential Body of Knowledge (EBK)	US Geospatial Intelligence Foundation	Competencies, Topics, Subtopics, Learning Objectives	March 2020
Geospatial Technology Competency Model (GTCM)	GeoTech Center	Competencies, Occupations, Job descriptions (including Learning Objectives)	March 2018
Geospatial Management Competency Model (GMCM)	Urban and Regional Information Systems Association	Critical work functions, Competency areas, Learning Objectives	June 2012
Developing A Curriculum (DACUM)	GeoTech Center	Learning Objectives	October 2019

- The Essential Body of Knowledge (EBK) came from the USGIF (Johnson 2019). Its content is limited only to topic titles and learning objectives, but it differentiates its learning objectives to hierarchical proficiency levels. It has five competencies (first level), 80 topics (second level), 946 subtopics (third level), 1285 learning objectives (fourth level). Subtopics in EBK are also classified as four different proficiency levels (Prerequisites, Foundational, Application, and Mastery).
- The Geospatial Technology Competency Model (GTCM) (DiBiase *et al*. 2010). The GTCM's general academic and workplace competencies (e.g the ability to think critically or to plan and organize) are outside the scope of our GIS-focused research, but its 'Industry-Wide Technical Competencies' and 'Industry-Sector Technical Competencies' are learning objectives specifically related to GIS. It has five competencies at the first level, and then competencies are associated with single or multiple occupations (second level). The detailed page for each occupation contains the learning objective (third level) of that occupation.
- The Geospatial Management Competency Model (GMCM) (Babinski 2012) is a derivative of the Geospatial Technology Competency Model. It exists in a matrix form including 74 rows corresponding to the critical work functions that most geospatial managers need to be able to perform and 18 columns that correspond to competency areas. This matrix indicates the associations between competencies and work functions.
- The Developing A Curriculum (DACUM) Job Analysis (Johnson 2010). During this process organized by the GeoTech Center, electronic surveys and in-person panel sessions were conducted to query a large number of GIS professionals about their daily work activities and practices. The results became the learning objectives that form the basis of a curriculum for future GIS professionals. Recently, the results from individual DACUM panels spanning the years 2008–2018 were collated and then mathematically ranked using regression analysis to arrive at a final DACUM.

While not explicitly labelled as such, these competency collections from the GTCM, GMCM and DACUM are similar to a 'body of knowledge' for professionals working in the GIS domain. Learning objectives and competencies or tasks are not identical in form or intent, but their generally similar purpose was adequate for inclusion in this ontology, particularly because we were aiming to cover both academic and professional GIS knowledge across the broad domain.

4.2. Reconciling the sources

Considering these sources as knowledge collections that could be integrated or merged was fundamental for our ontological work. Building a worthwhile ontology from scratch is laborious work that can be accomplished successfully with contributions from domain experts. In this situation, the knowledge already existed within these collections, produced via the earlier collaborative efforts by GIS experts. The sources are well defined and known to the GIS community. We were able to carefully compare these collections and identify the components that the sources had in common, such as learning objectives and competencies (Table 1). Their existence across our sources allowed them to become appropriate anchor points for our ontology engineering and facilitated our ranking of

the relatedness of the collections as we conducted the merging. Finally, a manual check was performed to ensure our ontology is complete and consistent. This workflow is presented in Figure 3. To organize our ontology systematically, we used a 4-level hierarchical tree structure based on the UCGIS GIS&T BoK. The top (root) level is the ten current knowledge areas: *Foundational Concepts, Knowledge Economy, Computing Platforms, Programming and Development, Data Capture, Data Management, Analytics and Modeling, Cartography and Visualization, and Domain Applications*. Every knowledge area has a short description. The second level is *unit* and the third level is *topic*. We denote the learning objective as the fourth level, because learning objectives exist in every topic and indicate the topic's focus. See Figure 2 for examples.

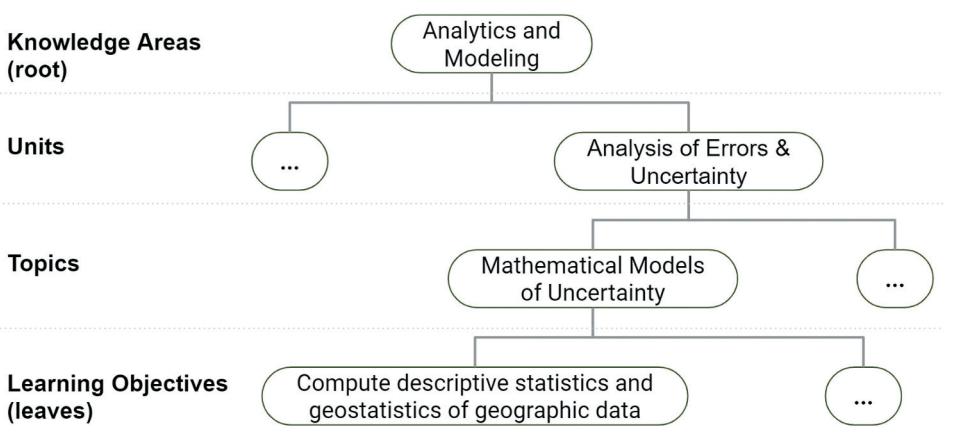

Figure 2. Our ontology examples. This graph shows an example path from knowledge area (root) to learning objectives (leaves).

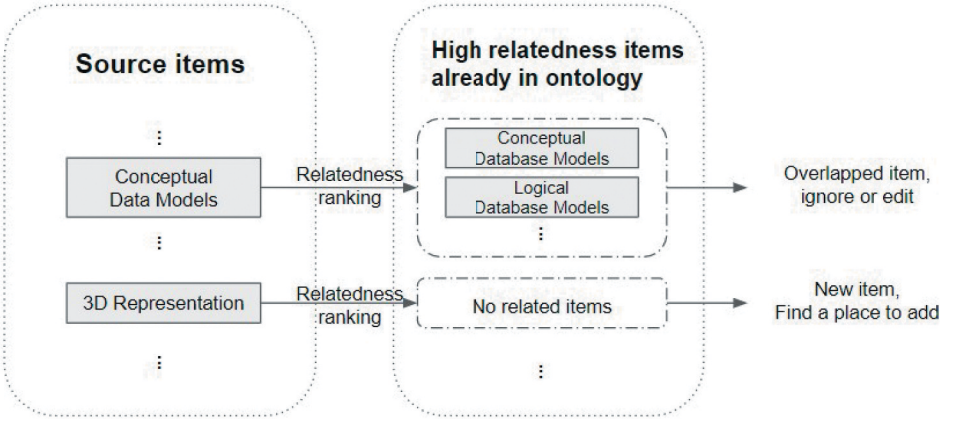

Figure 3. Workflow of ontology merging.

4.3. Relatedness ranking

To merge these collections, we found their similar entities, and inferred related entities that need to be added or deleted in our new ontology.

We first defined simple representing text (or SRT) and extended representing text (or ERT). SRT is the text used to describe the ontology entity itself. ERT is an extension of SRT that contains the graph structural information for adding the related entity of SRT.

Using UCGIS GIS&T BoK as an example, the simple representation for a knowledge area is its name and short description. For the other levels, the text identifying the unit, topic and learning objectives are their SRTs. The ERT is composed of a set of synonyms of the SRT based on Wordnet (Miller 1998) plus the SRTs of adjacent levels. For a fourth-level node (learning objective), the ERT is the entirety of the SRT in the path from the first level (knowledge area) down to the learning objective. A single topic may include several learning objectives, so we define the ERT for a topic as the SRT of all its learning objectives.

For instance, a third-level SRT is '*Fuzzy Aggregation Operators*' plus the synonyms '*"fuzzed", "fuzzy", "bleary", "blurred", "blurry", "foggy", "hazy", "muzzy", "aggregation", "accumulation", "assemblage", "collection", "collecting", "assembling", "aggregation", "manipulator", "operator"*'. The corresponding ERT would be the text in second and fourth level '*Problems of Scale and Zoning*' and '*Compare and contrast Boolean and fuzzy logical operations; Compare and contrast several operators for fuzzy aggregation, including those for intersect and union; Exemplify one use of fuzzy aggregation operators; Describe how an approach to map overlay analysis might be different if region boundaries were fuzzy rather than crisp; Describe fuzzy aggregation operators*'

We use h_s^l and h_e^l to denote the representation of a learning objective (l)'s simple representation text (SRT) and extended representation text (ERT), h_s^t and h_e^t for a topic(t)'s SRT and ERT, h_s^u for unit's SRT, h_s^k for knowledge area(k)'s SRT. Equation 1 and 2 show the calculation of a topic ERT and learning objective ERT calculation.

$$h_e^l = h_s^l + h_s^t + h_s^u + h_s^k \tag{1}$$

$$h_e^t = \sum_{i \in T} h_s^l(i) + h_s^t \tag{2}$$

Next, bag-of-words embedding was extracted from the SRT and ERT. Bag-of-words embedding builds an n-dimensional dictionary (n is the number of words) and represents text based on the word appearance in the dictionary. Stemming is used to remove any inflectional affixes in words (Manning *et al*. 2008). For example, 'discover' is the stemmed form of 'discovering', 'discovered' and 'discovery'. To only capture the key information, we stemmed all the text. We also removed the stop words such as 'a', 'the', and 'is' (see (Bird *et al*. 2009) for the full list of stop words). The bag-of-words embedding h used in Equations 1 and 2 are normalized by term frequency–inverse document frequency shown in Equation 3:

$$\mathrm{tf} - \mathrm{idf}(\mathrm{t}, \mathrm{d}) = \mathrm{tf}(\mathrm{t}, \mathrm{d}) \times \mathrm{idf}(\mathrm{t}) \tag{3}$$

in which term frequency (*tf*) is the number of times a term (*t*) occurs in a given document (*d*). The inverse document frequency (idf) is computed as

$$\mathrm{idf}(t) = \log \frac{1+n}{1+\mathrm{df}(t)} + 1 \tag{4}$$

where n is the total number of SRT, and $\mathrm{df}(i)$ is the number of SRT that contain term t.

The relatedness is measured by the cosine similarity between different topics' bag-of-words embedding of ERT h_e^t. Similarly, the learning objectives relatedness is defined as the cosine similarity between different bag-of-words embedding of ERT. Assuming the bag-of-words embedding representation of ERT is an n-dimensional vector, the cosine similarity between 2 ERT is defined as:

$$\cos(h_e(a), h_e(b)) = \frac{\mathbf{ab}}{\| \mathbf{a} \| \| \mathbf{b} \|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \sqrt{\sum_{i=1}^{n} (b_i)^2}} \tag{5}$$

The relatedness score can be used to measure the similarities across our source collections and can be used to recommend related topics or learning objectives to readers. It also shows the relationship between these different GIS collections of knowledge. After we examined all possible ERT pairs, we found that some topics and learning objectives demonstrated low relatedness, indicating they were independent or perhaps new knowledge that could be added to the ontology. The relatedness score is also useful in identifying when topics have similarities that might otherwise go unobserved due to their respective locations in the hierarchy.

Not surprisingly, the relatedness measurement revealed that these existing GIS bodies of knowledge and other collections overlap substantially but also have differences. We found 45 topics elsewhere that were not present in the UCGIS GIS&T BoK in its Q1 2020 version. Though the relatedness ranking automatically identifies the overlap among different sources and indicates the independent components, it does not have the ability to generate a new ontology and ensure that there is logical coherence for added topics, so a 'human-in-the-loop' is necessary here. We and other GIS experts at our universities manually considered each of these topics to decide whether they and their learning objectives should be added to the ontology. We followed the subsumption principles: term x subsumes y if y occurs only in a subset of the documents that x occurs in (Nyerges *et al*. 2014). The low relatedness entities were added to the bottom level of the ontology and the parent nodes are identified by the relatedness. In the end, the semantic similarity process helped us identify and add 12 topics and 116 learning objectives to the derived ontology. Our merged hierarchical ontology has 10 knowledge areas, 96 units, 412 topics and 1583 learning objectives.

5. Collecting additional GIS knowledge

Knowledge sources come in a variety of formats (books, research papers, blog posts, videos, etc.). Among them, scientific publications are traditionally considered to be a primary and authoritative source of knowledge. In addition to the text itself, a paper's authors, citations, and other details represents additional relevant data for researching collaborations, networks, and knowledge productivity. To further refine and evaluate our new ontology, we crawled a large collection of research papers from two very large open sources. In this section, we describe how we collected and processed those papers.

Table 2. Paper data set schema (Wang *et al*. 2019). Underlined fields are common field used to link different categories.

Category	Key information
Paper	PaperID, Title, Keywords, Abstract, Citations, References, VenueID, AuthorID, Mentioned locations
Venue	VenueID, Name, Publisher, Webpage, CreatedDate
Author	AuthorId, Name, LastKnownAffiliationId, CreatedDate
Affiliations	AffiliationId, Name, OfficialPage, CreatedDate

We relied on Open Academic Graph and Microsoft Academic Graph as the sources for the research publications (Sinha *et al*. 2015, Wang *et al*. 2019). As of 27 March 2020, their combined collection included a total of 234,049,193 publications, with more than 200 million individual authors and more than 50 thousand publication venues (every conference or journal volume counts as a venue) in the database. Words within our new ontology (see Section 4) became keywords to filter out only the GIS-related papers, resulting in 955,186 papers. Next, to ensure the quality of our collection, we filtered out any paper that had not already been cited at least once. English language publications were the major content in the original collections, though they do include other language publications if an English title is provided. Since we wanted to analyze more than a title alone, only English language papers were included in our analysis. In the end, we had identified 560,608 papers to be used for analysis, with 1,195,576 authors and affiliations (combined) and over 24,536 venues.

The data schema of the GIS knowledge collection is shown in Table 2. There are four major categories of information: the paper's meta information, plus associated venue, author, and author affiliation. They are connected through common fields.

6. Fusing the GIS knowledge

The large collection of GIS research papers (described in Section 5) was connected to our merged hierarchical ontology (described in Section 4) using deep learning-based methods to create the final GIS knowledge graph (GIS-KG). These may seem like disparate items, but they are deeply connected conceptually. An ontology defines what a domain includes and serves to distinguish the GIS research paper collection from a general academic corpus. While the papers are the source of concepts, they also naturally define the concepts' relations. Papers may explicitly state the relations between concepts, or they may simply indicate such relations through citations in their respective literature reviews. Once connected, presenting such linkages to users of the knowledge graph in a clear and unconfusing way is key. The match process was a multi-label classification task, which means each paper could be matched to one or more different entities in the ontology. Uncovering such linkages and making them discoverable and visible to the user community is a significant contribution to advancing the GIS knowledge domain.

6.1. Feature selection

Classifying papers requires that we know what features are available to be used for the task. Here, each research paper is used not for its text itself, but also its associated citation, references, and venues. Similar to how we represented ontology features, we constructed

simple representation text (SRT) and extended representation text (ERT) for the research papers' graph. For each research paper, the SRT is the text of its title, keywords, and abstract. A publishing venue's full name (i.e. the journal name or the conference name) is its SRT. We sampled a subset of publications from a given venue and concatenated their SRT. This was used as this venue's ERT. The ERT for publications included the SRT from its citations and references and the SRT of its linked publishing venue.

For example, the SRT of a subset of papers from the *International Journal of Geographic Information Science* (IJGIS) are used to construct the ERT for this venue. A publication's ERT includes the SRT from its citations and references and the ERT of its linked publishing venue. Then, we have

$$h_e^p = h_s^p + \sum_{i \in Cit} w_i h_s^p(i) + \sum_{j \in Ref} w_j h_s^p(j) + w_v h_e^v \tag{6}$$

$$h_e^v = \sum_{i \in V} h_s^p(i) + h_s^v \tag{7}$$

The h_s^p and h_e^p are the symbol of a publication p's SRT and ERT. h_s^v and h_e^v are a venue v's SRT and ERT. $\sum_{i \in Cit} h_s^p(i)$ were the papers that p cited. $\sum_{j \in Ref} h_s^p(j)$ were the papers where p was referenced. $\sum_{i \in V} h_s^p(i)$ were all the publications in that venue. Weight w is used to discount different neighbors' impact as appropriate. By inferring the relations between a paper and the ontology, all other information, like authors and venues, can be indirectly aggregated through the publications. Here, we use the $w = \frac{1}{n}$, (n is the number of citations or references, as a normalization term), so the number of citations and references of each paper will not have an impact on the representation.

6.2. Deep learning-based matching

To fuse the hierarchical ontology and knowledge materials together, we designed a deep learning-based matching rule. The matching process compared the similarity of a paper and the ontology. However, as both of these are unstructured text, it was difficult to directly compute the similarity. Thus, we used feature embedding to convert the unstructured data to vectors that can be calculated. We utilized a Siamese network, an embedding structure that uses a deep neural network to embed two kinds of objects in a consistent manner, to embed the ontology and publications and compare the similarity across embedding approaches (Figure 4) .

In our Siamese networks, the three types of embedding vectors were extracted bag-of-words, embedding-of-words, and embedding-of-sentences.

- **bag-of-words embedding**. The bag-of-words embedding used the same dictionary and techniques as the ontology. We used the same dictionary built as described in Section 4 and normalized by term frequency–inverse document frequency. This bag-of-words embedding matches exact keywords. It also evaluates each keyword's importance by determining how many times the word appeared in each document and in how many documents the keyword appeared.

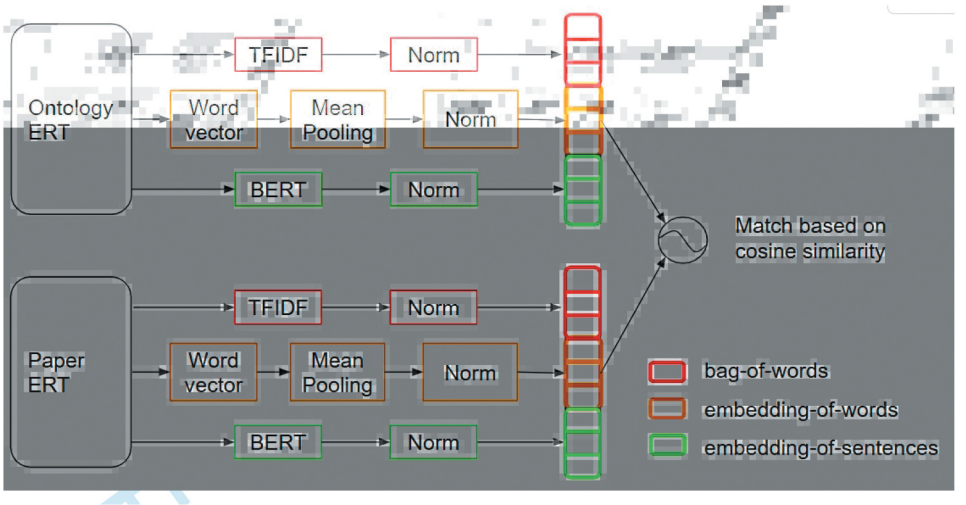

Figure 4. Deep learning-based matching structure. Ontology and publications use the same embedding method. ERT stands for the extended representation text.

- **embedding-of-words**. We trained a word embedding by using the skip-gram (Bojanowski *et al*. 2017) on the GIS research articles, with titles and abstracts. The resulting embedding-of-words model can generate 100-dimensional vectors for each word in the text. We calculated the mean value for all word vectors to represent the ontology and publications. The embedding-of-words took synonyms into account, so keywords with similar meanings would be matched.
- **embedding-of-sentences**. We fine-tuned the BERT model (Devlin *et al*. 2018) with our GIS research papers. This means we took the deep learning model structure and the original parameters in BERT, and continued the model training using our GIS research papers. The parameters were updated after our training, so the model was adapted to the GIS domain. The embedding-of-sentences model considered more text as contextual information in GIS while preserving its knowledge in general languages.

The three types of embedding were normalized and concatenated to achieve the final embedding for the ontology and publications. By combining the multiple embedding approaches, we produced a comprehensive representation of the ontology and knowledge materials. These features were concatenated for the vector representation h used in Equations 6 and 7. In the last step, we matched the publications to the ontology. The confidence score of an ontology-publication pair is the cosine similarity between these vector representations, a standard approach introduced by Wang *et al*. (2019).

In this way, we derived the final GIS-KG (the combination of the ontology and the research paper collection), which we could then evaluate for its effectiveness at information retrieval.

7. Knowledge retrieval as an application of our GIS-KG

The GIS-KG organizes GIS knowledge in a holistic knowledge graph, which has many benefits for the GIS community. For example, we tested its use in *ad hoc* searches for GIS publications where, given a query, a search engine would return several candidates that satisfy the query. Our experiment aimed to demonstrate that the GIS-KG could improve the search performance for both neural and non-neural information retrieval models. In this section, we compare the results of searching for GIS papers via the ontologically informed GIS-KG with searches for papers directly from the unstructured collection of GIS articles (as described in Section 5).

7.1. Analysis procedure

Environment settings. All experiments were run on a server that has 16 cores, 192GB RAM and five GPUs.

We ran the experiments in two different sets. The first set used existing retrieval models on the GIS paper collection. The second set is marked with '+*G*', which incorporated our GIS-KG that has the hierarchical ontology structure.

In the '+*G*' methods, we first used BERT to embed the query and simple representation of node for each node in the knowledge graph. Then, we computed the cosine similarity between the embedding of query and every node and get the top or first node for the query. Finally, we obtained the publications associated with the node and ran each methods on this subset of the publications.

- Both of the experiment sets used lexical matching methods and latent semantic models:
- *BM25* (Robertson and Zaragoza 2009). BM25 is a classic information retrieval algorithm based on bag-of-words representation. It is the default algorithm in Apache Lucene (also in Solr and Elastic search). We calculated the BM25 using the implementation based on Apache Lucene and we chose all the default parameters.
- *BM25 + G*. In this experiment, we first obtained the subset of publications based on our GIS academic graph. Then we used the same BM25 algorithm to retrieve publications.
- *WMD*(Kusner *et al*. 2015). Word Mover Distance (WMD) defines the similarity as the Wasserstein distance between word embeddings. We used the gensim implementation (Pele and Werman 2008, 2009) to calculate the similarity between queries and documents.
- *WMD+G*. We used the same WMD model with the subset of publications based on our GIS Academic Graph.
- *NWT*(Guo *et al*. 2016). Non-linear Word Transportation (NWT) is a deep learning-based model for information retrieval. It uses the maximum likelihood to combine both the exact match method and the inexact match method. This model can capture the semantic meanings because of its deep learning-based nature.
- *NWT+G*. We used the same NWT model as described above, with the difference that we retrieved the publications in the subset based on our GIS-KG.

- *BERT (*Devlin *et al*. 2018). The general pre-trained BERT model was used to embed the search query. The document was embedded with the same BERT model. Then, we calculated the cosine similarity between the embedding of search query and the embedding of documents.
- *BERT+G*. Our fine-tuned BERT model is used to embed text in this method. First, we compared the embedding similarity between the search query and the ontology ERT mentioned in Section 6, so we can get the ontology nodes that are related to the search query. Then, we retrieve the subset of publications from our GIS-KG based on the ontology. Finally, we embed the papers in the subset and compare them with the embedding of search query.
- One key step is to compare the similarity of queries and documents. We form this process as matrix multiplication:

$$\begin{aligned}&[\,simScore(q,d_1) \quad simScore(q,d_2) \quad \ldots \quad simScore(q,d_n)\,] = \\ &Embedding(q) \times [\,Embedding(d_1) \quad Embedding(d_2) \quad \ldots \quad Embedding(d_n)\,]_{m\times n}\end{aligned} \tag{8}$$

in which *simScore* is a scalar for the match score of a query (q) and a GIS article (d), the $Embedding(q)$ and $Embedding(d)$ are m dimensional embedding vectors, n is the number of GIS articles. We pre-compute the embedding of our ontology and publication graph $Embedding(D)$, so only the user query $Embedding(q)$ needs to be computed in real time. Then, the retrieved results can be sorted based on *simScore* with little cost. This enables the low latency of the online service to give immediate responses to a search.

7.2. Evaluation procedures and metrics

For each model, we conducted 50 queries and presented the retrieved results to human testers. The queries were generated from interviews with GIS experts by asking them *'what query do you use when searching for scientific publications?'*. The results were evaluated by graduate students and undergraduate students majoring in GIS at our universities, as well as GIS professionals through various social media platforms. We stopped the surveys once we had an adequate number of labelled results defined later in this section. The experts were asked to give scores between 0 and 4 to the retrieved documents using a simple ranking system.

As an example, a search for 'Spatial Cloud Computing' in a search engine would generate a list of papers in the results that could be ranked as indicated below:

- The paper *"Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?"*(Yang *et al*. 2011) and other papers that directly introduce the concept (including Spatial Grid/Parallel computing) would be highly relevant. These would be assigned a score of 4.
- A paper about the parallel algorithm for spatial data, or a web GIS system design paper, would be less directly relevant. These would be assigned a score of 3.
- A paper about an application of cloud computing with spatial data in one sentence is treated as borderline. This would be assigned a score of 2.
- A paper about a general cloud computing platform is not relevant. This would be assigned a score of 1.

- A paper talking about a cloud in the sky is considered not related. This would be assigned a score of 0.
- If the reviewer was uncertain, the result was scored a − 1 to flag it.

Evaluators used a general search engines (Microsoft Academic) to search for and review the content so its relevance can be clearly measured.

We used the majority vote to evaluate the results, and at least two people evaluated each result. If a − 1 value appeared, the result was assigned to another judge until the evaluation value fell between 0 and 4. If the absolute difference between evaluation values was less than 2, we took the average of the scores. If that difference was greater than 2, we asked a third person to evaluate, and took the third person's result as final.

It is not intuitive to compare all the evaluation scores directly, so four commonly used information retrieval algorithm evaluation metrics (Precision, Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR)) were utilized (Järvelin and Kekäläinen 2002, Robertson *et al*. 2010). The Precision@10 shows how many relevant results are in the top 10. The MAP is a cumulative way to calculate the precision, which represents the average precision in top one, top two, top three of the results, etc. The MRR considers the rank order of the relevant results. For example, if the first result is relevant, then the MRR score is 1. If the first result is irrelevant but the second result is relevant, then the MRR score is 1/2. The NDCG shows the order prediction ability for different models. For different models, with a higher score indicating that a model is better at ranking the results. A better NDCG score assumes that the higher relevance results should always rank higher than those lower relevance results. We calculated these metrics on our human evaluation results (Table 3).

7.3. Experiment results

From the experimental results shown in Table 3, we can see that:

- **GIS-KG is useful**. In this information retrieval task, all baseline models (those without "+*G*") improved from 4.2% to 73.6% with the addition of graph enhanced methods. Precision scores represent how many relevant or related articles are present in the results, and the BM25 returned only about 50% of the results relevant to the query. But when we used the GIS-KG, more than 80% of the results were relevant. This BM25 +G excelled over one deep learning-based baseline method (the WMD). The NDCG

Table 3. Evaluation results of the information retrieval task. The comparison is between the corresponding +*G* and without the G methods. The greater the number, the better the results.

Experiment	Precision@10	NDCG	MAP	MRR
BM25	0.555	0.609	0.330	0.559
BM25 + G	0.827(↑ 49.0%)	0.809(↑ 32.8%)	0.573(↑ 73.6%)	0.818(↑ 46.3%)
WMD	0.706	0.700	0.467	0.700
WMD+G	0.790(↑ 11.9%)	0.831(↑ 18.7%)	0.550(↑ 17.8%)	0.810(↑ 15.7%)
NWT	0.882	0.849	0.610	0.932
NWT+G	0.891(↑ 1.0%)	0.894(↑ 5.9%)	0.646(↑ 6.0%)	1(↑ 7.3%)
BERT	0.908	0.858	0.635	0.935
BERT+G	0.954(↑ 5.0%)	0.878(↑ 2.3%)	0.673(↑ 6.0%)	0.975(↑ 4.2%)

shows the order prediction ability for different models. Higher scores mean a model can better rank the results in the right order. Results of MAP and MRR also present consistent findings of other indicators.

- **GIS-KG boosts traditional methods more than advanced methods**. The GIS-KG helps substantially when compared to the most traditional methods (BM25). It also improved the performance for more recent baselines, though with a narrower margin. The BM25 model can only retrieve results that have the same keywords as the query. This exact match method would miss many semantically similar results and would also return matches having the same keywords with different meanings. The GIS-KG linked papers with related concepts together, enabling the model to discover more related papers. The GIS-KG also sets constraints so that only the papers with associated concepts can be matched by the algorithm. Compared with a general language model that can understand semantic meanings, the GIS-KG contains domain expert's knowledge. This domain knowledge is the key to better search results even when compared to the state-of-the-art language models. Although the recent information retrieval models already perform relatively well, small improvements are still highly valuable and worthwhile.
- **BERT+G is the best method in our setting**. The scores in Table 3 shows NWT+G and BERT+G generally outperformed other methods. The NWT+G had higher NDCG and MAP scores than the BERT+G method. This means the NWT+G can better sort the relevance of the retrieved papers. The Precision and MAP scores are higher in the BERT+G method, which means the BERT+G method found more related documents than the NWT+G method. When conducting information retrieval manually, we would prioritize the overall number of related results, so we consider the BERT+G model to be optimal.

8. Discussion

The quantitative and qualitative studies have demonstrated the effectiveness of our GIS-KG for search queries on a GIS-focused collection. Our work can be further improved as follows:

- **More complete source materials**. The UCGIS GIS&T Body of Knowledge is undergoing continuous expansion. Even since the Q1 2020 version that we used for our ontology, several dozen topics and their respective learning objectives have been added to the collection. The European's EO4GEO has also become available as open source since our research was completed. Integration of these expanded sources will continue to refine our ontology.
- **More abundant knowledge materials**. This work focused on textual data, but the GIS-KG can be extended to other formats of knowledge materials. For example, images in publications usually carry rich information. Numerous texts and video-based tutorials exist on the Internet that are designed to teach students how to operate GIS software and practice fieldwork. Even the code of GIS software is a type of knowledge material. Using a concept similar to GIS-KG, we can identify and then organize access to images, videos, and software that are used in the practice of teaching and learning GIS.

- **Better search quality**. Our experimental results already indicate that we can achieve better information retrieval results with more advanced information retrieval algorithms. It is important to note that we cannot directly compare the process and results from our whole system with Google Scholar because 1) the dataset we used is different from what Google crawls from the web; 2) the search mechanism deployed by Google is unknown to us, and it might change anytime as Google is constantly upgrading their algorithms; and 3) Google collects personal information to advance and refine their search results, thus undermining searches being evaluated systematically and scientifically. That means that searches are likely to return different results by different people at different times on different computers. There are too many variables we cannot control. Instead, in this paper, we conducted a controlled experiment. We kept the dataset and algorithm constant and only varied whether GIS-KG was used or not. As a result, our results are more certain and convincing because the dataset and algorithm are independent. More advanced retrieval models and algorithms will further improve search results. When additional knowledge materials exist within the collection, users will be able to find their most desired or appropriate type of knowledge content: research papers, video tutorials, etc. This could be achieved with the design of advanced algorithms and deep understanding of the domain practitioner's needs.
- **Stronger graph knowledge discovery**. The rich graph information was utilized when we built the graph, but it was not used when we conducted the information retrieval experiments. The relationship among the entities is another type of valuable knowledge that remains to be explored. It contains the logical flow and development information of a knowledge domain. Constructing such links across knowledge sources is evidence for having mastered the knowledge itself. Graph reasoning and link prediction algorithms may be applied to this GIS-KG to study the structure of the GIS-KG and derive new knowledge from it. We leave it for further study.

9. Conclusion

The premise of this paper is that a systematically organized knowledge base is critical for domain experts to conduct scientific research, academic education, technical training, and professional practice. As demonstrated in this paper, we significantly advanced the vision for a comprehensive GIS&T Body of Knowledge by merging multiple collections into a single ontology and enhanced the tool further by linking a collection of scientific publications to that ontology. We built a large-scale hierarchical geographic information science and technology knowledge graph, GIS-KG, in which over 500 thousand publications are organized within an integrated GIS topic structure based on deep learning methods. At the center of our contribution lies an AI-assisted ontology re-design for the UCGIS GIS&T BoK, one that brings together various knowledge sources in a comprehensive manner, using existing experts' wisdom to organize knowledge efficiently and effectively. In the knowledge graph building process, we successfully migrated natural language processing technologies and adapted those to the GIS domain, informed by semantic relatedness in the GIS field.

We demonstrated the utility of this GIS-KG by conducting GIS information retrieval tasks. Using our GIS-KG, we improved retrieval quality in both traditional and deep learning methods. We provide open access to all the data and tools used in this work to help the broader domain community understand the structure and content of the GIS-KG itself and, more importantly, to help explore the knowledge within the abundant and rapidly growing context of scientific publications.

Going forward, this comprehensive knowledge graph will boost knowledge discovery in the GIS field. We seek cooperation and input from other GIS professional groups to advance the GIS-KG and continue the development of its framework. We encourage the GIS community to further explore our GIS-KG and build more applications on it. We believe this new knowledge graph will become a valuable resource for the community.

Acknowledgements

We greatly appreciate the helpful comments and suggestions from the editor and anonymous reviewers. The research was supported by National Science Foundation (NSF) under grants OIA-1937908 and SMA-2122054, Texas A&M University Harold Adams Interdisciplinary Professorship Research Fund, and College of Architecture Faculty Startup Fund. The funders had no role in the study design, data collection, analysis, or preparation of this article. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the NSF [1937908, 2122054]; Texas A&M University Harold Adams Interdisciplinary Professorship Research Fund; Texas A&M University College of Architecture Faculty Startup Fund

Notes on contributors

Jiaxin Du is a Ph.D. student at the Department of Landscape Architecture and Urban Planning at Texas A&M University. He holds M.S. and B.S. in Geographical Information Systems from Zhejiang University. His research focuses on geospatial artificial intelligence, natural language processing, and their application in urban planning.

Shaohua Wang is an assistant professor in the Department of Informatics, College of Computing, New Jersey Institute of Technology. His research interests include software engineering, program analysis, and artificial intelligence. Dr. Wang has published research on top computer science conferences and journals, such as ICSE, FSE, ASE, OOPSLA and TSC.

Xinyue Ye is a Harold Adams Endowed Associate Professor at the Department of Landscape Architecture and Urban Planning at Texas A&M University. He holds a Ph.D. degree in Geographic Information Science from the Joint Program between University of California at Santa Barbara and San Diego State University, a M.S. in Geographic Information Systems from Eastern Michigan University, and a M.A. in Human Geography from University of Wisconsin at Milwaukee. His research focuses on geospatial artificial intelligence, smart cities, spatial econometrics, and urban computing.

Diana S. Sinton focuses on the teaching and learning of geographic information science and systems (GIS), especially in the natural and environmental sciences. She serves as a Senior Research Fellow for the University Consortium for Geographic Information Science (UCGIS), a non-profit scientific and educational organization that supports a community of practice around GIScience research and teaching in higher education. Sinton teaches courses in GIS and spatial analysis at Cornell University in Ithaca, New York.

Karen Kemp is Professor Emerita of the Practice of Spatial Sciences in the Spatial Sciences Institute at the University of Southern California Dornsife College of Letters, Arts and Sciences.

Her scientific research has focused on developing methods to improve the integration of environmental models with GIS from both the pedagogic and scientific perspectives and on formalizing the conceptual models of space acquired by scientists and humanities scholars across a wide range of disciplines.

ORCID

Xinyue Ye http://orcid.org/0000-0001-8838-9476
Diana S. Sinton http://orcid.org/0000-0001-5828-9001

Data availability statement

The code and data that support the findings of this study are available from https://github.com/UrbanDS/GIS-KG.

References

Adhikari, A., *et al*., 2019. Docbert: bert for document classification. arXiv preprint arXiv:1904.08398.

Ahearn, S.C., *et al*., 2013. Re-engineering the gis&t body of knowledge. *International Journal of Geographical Information Science*, 27 (11), 2227–2245. doi:10.1080/13658816.2013.802324.

Al-Taha, K.K., Snodgrass, R.T., and Soo, M.D., 1994. Bibliography on spatiotemporal databases. *International Journal of Geographical Information Systems*, 8 (1), 95–103. doi:10.1080/02693799408901988

Babinski, G., 2012. Urisa develops the geospatial management competency model (gmcm) for usdoleta. In: Washington GIS Conference, Washington. vol. 9.

Biljecki, F., 2016. A scientometric analysis of selected giscience journals. *International Journal of Geographical Information Science*, 30 (7), 1302–1335. doi:10.1080/13658816.2015.1130831.

Bird, S., Klein, E., and Loper, E., 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. Sebastopol, California: O'Reilly Media, Inc.

Bojanowski, P., *et al*., 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. doi:10.1162/tacl.a.00051.

Bostock, M., Ogievetsky, V., and Heer, J., 2011. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17 (12), 2301–2309. doi:10.1109/TVCG.2011.185.

Bourque, P., *et al*., 2014. Guide to the software engineering body of knowledge (swebok) and the software engineering education knowledge (seek)-a preliminary mapping. *In: Proceedings 10th International Workshop on Software Technology and Engineering Practice*. IEEE Computer Society Press, 8–8 .

Brown, T.B., *et al*., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Cochrane, L., *et al*., 2017. Searching for social justice in giscience publications. *Cartography and Geographic Information Science*, 44 (6), 507–520. doi:10.1080/15230406.2016.1212673.

Dai, Z., and Callan, J., 2019. Deeper text understanding for ir with contextual neural language modeling. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval,* Paris, France. 985–988.

Devlin, J., *et al*., 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

DiBiase, D., *et al*. 2006. Introducing the first edition of geographic information science and technology body of knowledge. *Cartography and Geographic Information Science*, 34 (2), 113–120. doi:10.1559/152304007781002253

DiBiase, D., *et al*., 2010. The new geospatial technology competency model: bringing workforce needs into focus. *Journal of the Urban & Regional Information Systems Association*, 22 (2), 55.

Fan, W., *et al*., 2019. Deep social collaborative filtering. In: *Proceedings of the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, 305–313.

Fricke, S.N., 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106 (1), 145–147. doi:10.5195/JMLA.2018.280

Ghorbanzadeh, O., *et al*. 2020. A new gis-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. *Journal of Spatial Science*, 65 (3), 401–418. doi:10.1080/14498596.2018.1505564

Guo, J., *et al*., 2016. Semantic matching by non-linear word transportation for information retrieval. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, New York, NY, USA: Association for Computing Machinery, 701–710. doi:10.1145/2983323.2983768.

Gusenbauer, M., 2018. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118 (1), 177–214. doi:10.1007/s11192-018-2958-5

Haklay, M., Singleton, A., and Parker, C., 2008. Web mapping 2.0: the neogeography of the geoweb. *Geography Compass*, 2 (6), 2011–2039. doi:10.1111/j.1749-8198.2008.00167.x

Harzing, A.W. and Alakangas, S., 2017. Microsoft academic is one year old: the Phoenix is ready to leave the nest. *Scientometrics*, 112 (3), 1887–1894. doi:10.1007/s11192-017-2454-3

Herzig, J., *et al*., 2020. Tapas: weakly supervised table parsing via pre-training. In: ACL.

Hofer, B., *et al*., 2020. Complementing the european earth observation and geographic information body of knowledge with a business-oriented perspective. *Transactions in GIS*, 24 (3), doi:10.1111/tgis.12628

Janowicz, K., 2021. Knowwheregraph drives analytics and cross-domain knowledge. *ArcUser*, 24 (Preprint), 16–19.

Järvelin, K. and Kekäläinen, J., 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20 (4), 422–446. doi:10.1145/582415.582418

Johnson, A., 2019. *Quest to define the knowledge and skills needed by geospatial professionals*. Cham: Springer International Publishing. 271–288. doi:10.1007/978-3-030-04750-4_14

Johnson, J., 2010. What gis technicians do: a synthesis of dacum job analyses. *Journal of the Urban & Regional Information Systems Association*, 22 (2), 31.

Kusner, M., *et al*., 2015. From word embeddings to document distances. In: *International conference on machine learning*, Lille, France, 957–966.

Li, W., Song, M., and Tian, Y., 2019. An ontology-driven cyberinfrastructure for intelligent spatio-temporal question answering and open knowledge discovery. *ISPRS International Journal of Geo-Information*, 8 (11), 496. doi:10.3390/ijgi8110496

MacAvaney, S., *et al*., 2018. Characterizing question facets for complex answer retrieval. In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '18*, New York, NY, USA: Association for Computing Machinery, 1205–1208. doi:10.1145/3209978.3210135.

Mai, G., *et al*., 2020a. SE-KGE A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24 (3), doi:10.1111/tgis.12629

Mai, G., *et al*., 2020b. Multi-scale representation learning for spatial feature distributions using grid cells. In: *International Conference on Learning Representations.* Virtual Conference, Formerly Addis Ababa ETHIOPIA. Available from: https://openreview.net/forum?id=rJljdh4KDH

Malczewski, J. and Jankowski, P., 2020. Emerging trends and research frontiers in spatial multicriteria analysis. *International Journal of Geographical Information Science*, 34 (7), 1257–1282. doi:10.1080/13658816.2020.1712403

Manning, C.D., Schütze, H., and Raghavan, P., 2008. *Introduction to information retrieval*. Cambridge, England: Cambridge university press.

Marble, D., 2001. A bibliography on the application of geographic information science and geographic information technology to arctic problems. In: *ESRI User Conference* San Diego, California, USA, 1 .

Marble, D.F., 2000. *A bibliography on the application of giscience and gi technology to disease and public health problems*. Columbus, Ohio, USA: Center for Mapping, The Ohio State University.

Martín-Martín, A., *et al*. 2018. Google scholar, web of science, and scopus: a systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12 (4), 1160–1177. doi:10.1016/j.joi.2018.09.002

Miller, G.A., 1998. *Wordnet: an electronic lexical database*. Cambridge, Massachusetts, USA: MIT press.

Nyerges, T., *et al*., 2014. Foundations of sustainability information representation theory: spatial–temporal dynamics of sustainable systems. *International Journal of Geographical Information Science*, 28 (5), 1165–1185. doi:10.1080/13658816.2013.853304.

Pele, O., and Werman, M., 2008. A linear time histogram metric for improved sift matching. In: *European conference on computer vision* Marseille, France, Springer, 495–508.

Pele, O., and Werman, M., 2009. Fast and robust earth mover's distances. In: *2009 IEEE 12th International Conference on Computer Vision* Kyoto, Japan, IEEE, 460–467.

Robertson, S.E., Kanoulas, E., and Yilmaz, E., 2010. Extending average precision to graded relevance judgments. In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* Geneva, Switzerland, 603–610.

Robertson, S. and Zaragoza, H., 2009. The probabilistic relevance framework: bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3 (4), 333–389. doi:10.1561/1500000019.

Siabato, W., *et al*., 2014. Timebliography: a dynamic and online bibliography on temporal gis. *Transactions in GIS*, 18 (6), 799–816. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12080 .

Sinha, A., *et al*., 2015. An overview of microsoft academic service (MAS) and applications. In: *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web, may*, New York, New York, USA: Association for Computing Machinery, Inc, 243–246 Available from: http://dl.acm.org/citation.cfm?doid=2740908.2742839 .

Skupin, A., 2014. Making a mark: a computational and visual analysis of one researcher's intellectual domain. *International Journal of Geographical Information Science*, 28 (6), 1209–1232. doi:10.1080/13658816.2014.906040.

Tang, J., *et al*., 2008. Arnetminer: extraction and mining of academic social networks. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* Las Vegas, Nevada, USA, 990–998.

Tomaszewski, B., and Holden, E., 2012. The geographic information science and technology and information technology bodies of knowledge: an ontological alignment. In: *Proceedings of the 13th annual conference on Information technology education* Calgary Alberta Canada, 195–200.

Trisedya, B.D., Qi, J., and Zhang, R., 2019. Entity alignment between knowledge graphs using attribute embeddings. In: *Proceedings of the AAAI Conference on Artificial Intelligence* Hilton Hawaiian Village, Honolulu, Hawaii, USA, vol. 33, 297–304.

Vandenbroucke, D. and Vancauwenberghe, G., 2016. Towards a new body of knowledge for geographic information science and technology. *Micro Macro Mezzo Geoinf*, 6, 7–19.

Vaswani, A., *et al*., 2017. Attention is all you need. *In*: *Advances in neural information processing systems*, Long Beach Convention & Entertainment Center, Long Beach, CA, USA, 5998–6008.

Wang, C., *et al*., 2020. Digital earth education. *In*: *Manual of digital earth*, Singapore: Springer, 755–783.

Wang, K., *et al*., 2019. A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2, 45. Available from: https://www.frontiersin.org/article/10.3389/fdata.2019.00045/full

Wang, W., *et al*. 2021. Qa4gis: a novel approach learning to answer gis developer questions with api documentation. *Transactions in GIS*, 25 (5), 2675–2700. doi:10.1111/tgis.12798

Wikle, T. and Sinton, D., 2020. The administration of academic gis certificates: a survey of program coordinators. *Transactions in GIS*, 24 (6), 1681–1694. doi:10.1111/tgis.12677

Yang, C., *et al*. 2011. Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? *International Journal of Digital Earth*, 4 (4), 305–329. doi:10.1080/17538947.2011.587547

Ye, X., *et al*. 2021. Geospatial and semantic mapping platform for massive covid-19 scientific publication search. *Journal of Geovisualization and Spatial Analysis*, 5 (1), 1–12. doi:10.1007/s41651-021-00073-y

Zaharia, M., *et al*., 2016. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59 (11), 56–65. doi:10.1145/2934664.

Zalta, E.N., *et al*., 1995. Stanford encyclopedia of philosophy.

Zhan, F.B., Gong, X., and Liu, X., 2014. Mark on the globe: a quest for scientific bases of geographic information and its international influence. *International Journal of Geographical Information Science*, 28 (6), 1233–1245. doi:10.1080/13658816.2013.847186.

Zhong, S., *et al*., 2015. Progress in human geography in a century: a bibliometric review of 73 ssci journals. *Acta Geographica Sinica*, 70 (4), 678. pages 10 Available from http://www.geog.com.cn/EN/abstract/article36409.shtml