GAKG: A Multimodal Geoscience Academic Knowledge Graph

Cheng Deng[1], Yuting Jia[1], Hui Xu[1], Chong Zhang[1], Jingyao Tang[1], Luoyi Fu[1], Weinan Zhang[1], Haisong Zhang[2], Xinbing Wang[1], Chenghu Zhou[3]
[1]Shanghai Jiao Tong University, [2]Tencent AI Lab
[3]Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences
{davendw,hnxxjyt,yiluofu}@sjtu.edu.cn,hansonzhang@tencent.com,zhouch@lreis.ac.cn

ABSTRACT

The research of geoscience plays a strong role in helping people gain a better understanding of the Earth. To effectively represent the knowledge (KG) from enormous geoscience research papers, knowledge graphs can be a powerful means. However, the existing geoscience KGs mainly focus on the external connection between concepts, whereas the potential abundant information contained in the internal multimodal data of the paper is largely overlooked for more fine-grained knowledge mining. To this end, we propose **GAKG**, a large-scale multimodal academic KG based on 1.12 million papers published in various geoscience-related journals. In addition to the bibliometrics elements, we also extracted the internal illustrations, tables, and text information of the articles, and obtain the knowledge entities of the papers and the era and spatial attributes of the articles, coupling multimodal academic data and features. Specifically, GAKG realizes knowledge entity extraction under our proposed Human-In-the-Loop framework, the novelty of which is to combine the techniques of machine reading and information retrieval with manual annotation of geoscientists in the loop. Considering the fact that literature of geoscience often contains more abundant illustrations and time scale information compared with that of other disciplines, we extract all the geographical information and era from the geoscience papers' text and illustrations, mapping papers to the atlas and chronology. Based on GAKG, we build several knowledge discovery benchmarks for finding geoscience communities and predicting potential links. GAKG and its services have been made publicly available and user-friendly.[1]

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Semantic networks**; *Ontology engineering*; *Reasoning about belief and knowledge.*

KEYWORDS

Geoscience Multimodal Academic Knowledge Graph, Information Extraction, Data Management, Knowledge Base

[1] https://gakg.acemap.info/

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482003

ACM Reference Format:
Cheng Deng[1], Yuting Jia[1], Hui Xu[1], Chong Zhang[1], Jingyao Tang[1], Luoyi Fu[1], Weinan Zhang[1], Haisong Zhang[2], Xinbing Wang[1], Chenghu Zhou[3]. 2021. GAKG: A Multimodal Geoscience Academic Knowledge Graph. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482003

1 INTRODUCTION

Geoscience is a natural science that studies the earth, including Geography, Physics, Chemistry, and other disciplines [7]. Throughout history, paleontologists investigate the characteristics of various species and environmental evolution on the earth from 4.6 billion years ago to the present and explore the impact of environmental changes on biodiversity [9]. Geographers study topography, landforms, and climate, and find out that global warming caused by human activities has a certain relationship with the earth's axis drift [1]. Geologists explore the sea to bring more important resources such as rare earth minerals to mankind [13]. Thus it can be seen that geoscience plays an important role in the academic field, not only informing our understanding of the relationship between human beings and the earth but also helping us understand current change.

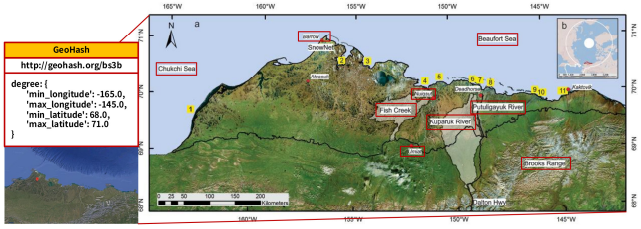


Figure 1: An Example of Illustrations in Geoscience Papers.

Academic papers have been employed as a major means to disseminate knowledge. Over time, scientists have published a large number of papers and accumulated a large knowledge system in the process of exploration and discovery of the earth. Distinguishing from other disciplines' papers, geoscience papers contain more abundant multimodal data such as geographical maps, tables, and era descriptions, reflecting the time and spatial characteristics. On one hand, different from other types of paper illustrations, maps can allow people to obtain spatial perception through visual cognition based on geographic information visualization. The maps in the geoscience papers' illustrations also have geographic location coordinates or atlas serving as small pieces of a world map (taking watersheds of the Alaska Arctic Coastal Plain [33] as an example in Figure 1, we can extract coordinates from it), so that the spatial

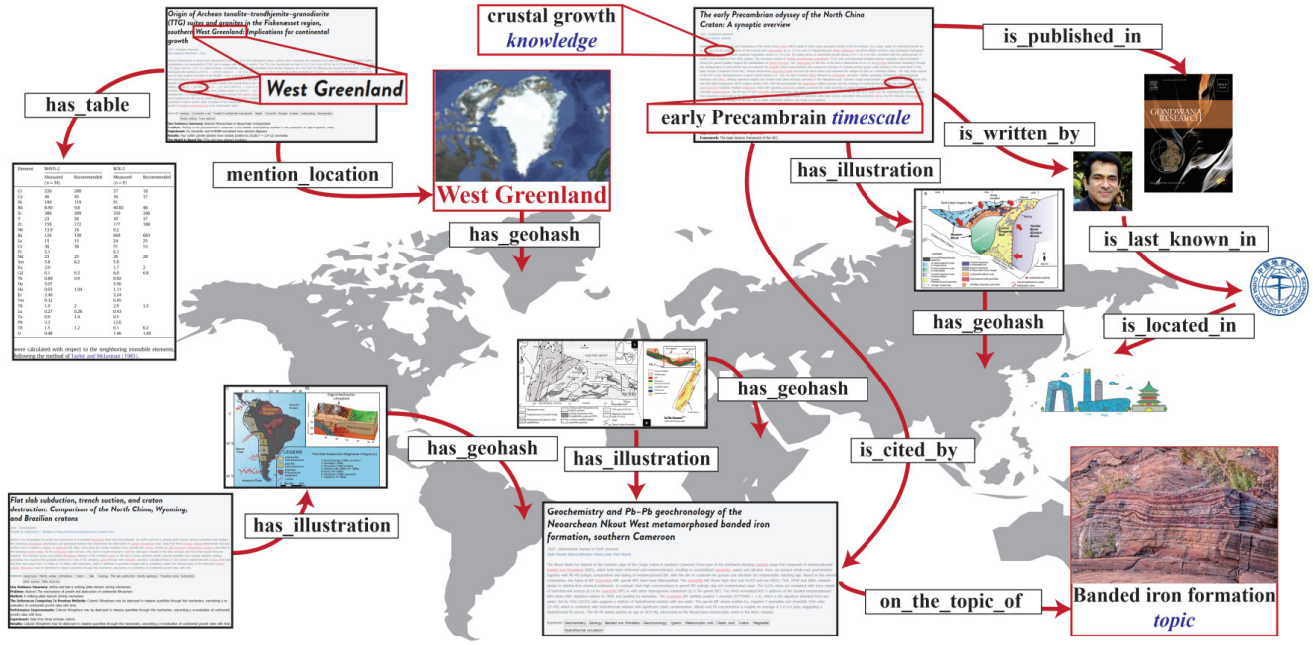


Figure 2: Overview of Multimodal GeoScience Academic Knowledge Graph (GAKG). @ Cheng Deng

relationship of these papers can be connected through the world map. Therefore, for geoscience, it is of great significance to collect and manage the papers' illustrations and mapping them to the geographic maps. On the other hand, geoscience papers not only share spatial information but also reveal geologic time scale (GTS) characteristics. For example, the GTS of [22] is early Precambrian instead of its publication time. Meanwhile, the locations mentioned in [35] include Brazil which is far away from the authors' affiliation. Hence, the era corresponding to papers may not be the publication time of the paper, and the geographical locations described in the papers may not be the residences of the scholars, showing the necessity to extract and manage the geographic locations and description GTS mentioned in the geoscience articles. As an extension, time scale information gives researchers a chance to manage, model, and analysis evolving knowledge. [17, 18]

Academic knowledge graphs (KGs) are usually an appropriate choice for integrating all the papers and related elements like authors and affiliations for specific disciplines. The existing academic KGs are mostly data collections based on bibliometrics, with typical examples belonging to AceKG [37], MAG [30], and AMiner Knowledge Graph [34]. These three academic KGs are mainly built by articles in English of all the scientific fields. This type of academic KG consists of bibliometrics entities such as authors, journals, papers, etc., linked through a few relations. For geosciences, academic-related knowledge graphs are rare. The existing KGs related to geoscience are mainly concept-level KGs that each entity represents a geographic object. The Common Sense Geographic Knowledge Base (CSGKB) [42] contains the edges that link the concepts of geographic features, geographic locations, and spatial relations. GeoKG [38] is a conceptual and formal geographic knowledge representation model based on geographic entities and their six elements such as state, location, change, attribute, time, and relation. Since, distinguishing from other disciplines, the academic data of geosciences are rich in geographic location, geologic time scale, and geographical maps reflecting the multimodality in geoscience. Bibliometrics information is not comprehensive towards geoscience academic data, and the discrete images, texts, and time scale within articles are not sufficiently coupled. As a whole, geoscience KGs [44] break new ground where geoscience, computer science, and information science converge. To this end, it is essential to have a job to manage multimodal geoscience knowledge, including scholars' elements and the knowledge, time scale information, and illustrations in their papers. Thus we can use the map as the building blocks to construct the geoscience academic knowledge graph, and comb intellectual dot of geoscience to build the knowledge system for this discipline. When it comes to using KG to organize multimodal academic data, Deep Code Curator project [12] puts forward a multimodal KG for deep learning with the code and illustrations of papers, and COVID-KG [36] extracts proper nouns and their illustrations. However, the information delivered by their illustrations has less dimension compared to the geoscience papers' images.

As a core concept in the academic KG, paper shapes the body of the academics world, and the internal knowledge of each paper is the soul of academics. Thus, it is also of significance to mining the inner knowledge of articles to expand academic KG. At present, the works on extracting the internal information of research articles to enrich the KG mainly focus on medicine. COVID-KG also gathers all the papers on the COVID-19 pandemic and process entity extraction on genes, diseases, chemistry, and biology entities in the content of the articles linking them to relevant external medical ontology. However, this requires a lot of manual participation to perform named entity annotation, and it may consume more resources when expanding to a larger academic field. Thus, minimizing manpower becomes the key issue to mine academic knowledge mining with high-precision requirements.

Contributions. As we will describe in the following sections in details, we design a novel multimodal academic knowledge graph for geoscience by collecting geoscience bibliometrics data and extracting each paper's illustrations, geographical information, geologic time scale, and inner knowledge entities, with Figure 2 illustrating the overview of GAKG and detailed contributions are as follows:

- We propose a multimodal **GeoScience Academic Knowledge Graph (GAKG)** framework by fusing papers' images, text, and bibliometric data. With multimodality, GAKG shed light on a new perspective for academic data mining and the construction of academic knowledge graphs and enriches the diversity of academic information retrieval.
- With a geographic world map, all the illustrations, text, and geologic time scale extracted from the selected geoscience papers construct a strong correlation and high coupling relationship between papers. In this way, scientists in the field of geology, geography, and data mining can carry out rich scientific research based on geographic locations.
- We put forward a Human-In-the-Loop knowledge extraction pipeline to extract paper's knowledge entities and mapping them to a crowd-sourcing knowledge taxonomy. In this way, we minimize human efforts and increase the precision of knowledge extraction using human-computer interactive annotation.
- To our knowledge, GAKG is currently the largest and most comprehensive geoscience academic knowledge graph, consisting more than 68 million triples. In order to better serve the data mining and knowledge discovery communities, GAKG is updated regularly that can be queried at SPARQL query Endpoint and explored at online applications.

The rest of the paper is organized as follows. In Section 2, we list the ontologies of GAKG and share the currently dumped datasets. In Section 3, the pipeline of multimodal GAKG is introduced. Based on GAKG, benchmarks for community detection and link prediction will be presented in Section 4. Moreover, we share the online applications in Section 5. Finally, the observations and the related works towards GAKG will be discussed in Sections 6 and 7.

2 OVERVIEW OF GAKG

As mentioned earlier, GAKG is a large-scale multimodal academic KG, with all the data collected from AceMap (https://www.acemap.info/). In this section, we will introduce the ontology and schema of the GAKG, followed by the current statistics of GAKG datasets. GAKG is updated regularly in accordance with its ontology.

2.1 GAKG Ontology

GAKG's schema-graph consists of **11** concepts connected by **19** relations. Five of them (*has_concluded, has_designed, is_located_in, has_developed and earn_in_the_way_of*) have a upper class relation *acer:mention_knowledge*. Since GAKG is the union of academic concepts and their relations, we manage GAKG as linked open data (LOD), we provide *#sameAs* axioms linking to the entities in other datasets.

The Graph base namespace (Graph IRI) is https://https://www.acekg.cn, all the concepts and relations shared. GAKG defines 11 classes, 19 object properties, and 39 data properties. The PREFIX set is shown in Figure 4(a).

Concepts. Each entity has a class type, a highly abstracted concept. The concepts we design can be listed as follows:

Paper (ace:paper) Representation of the academic papers in the field of geoscience. Concept *ace:paper* has 10 data properties including title(label), abstract, DOI, original URL, year and date the paper is published, ISSUE, volume as well as the start page and the end page of the journal. Among above, property title reuses the axiom *rdfs:label* and property original URL reuses axiom *foaf:page*.

Journal (ace:journal) Representation of academic journals in the field of geoscience. Concept *ace:journal* has **3** data properties including normalized name (reusing axiom *rdfs:label*), url (reusing axiom *foaf:homepage*) and ISSN.

Author (ace:author) Representation of the scholars in the field of geoscience who have published research articles in the 194 journals we selected. Concept *author* has **2** data properties including author's name (reusing axiom *rdfs:label*) and a date that the author published his/her last manuscript.

Affiliation (ace:affiliation) Representation of the affiliations where authors in the field of geoscience work in. Concept *ace:affiliation* has **5** data properties including its name (reusing axiom *rdfs:label*), abbreviation, homepage (reusing axiom *foaf:homepage*), its grid code and its introduction.

Topic (ace:topic) Representation of the academic topics of geoscience. The AceMap system tag each paper with the key phrase, we integrate them and establish the concept topic. Concept *ace:topic* has **3** data properties including topic name (reusing axiom *skos:prefLabel*) and definition (reusing axiom *skos:definition*) and a related image url.

Illustration (ace:illustration) Representation of the pictures in papers in the field of geoscience. Concept *illustration* has **3** data properties including illustration's tag, caption and dpi.

Papertable (ace:papertable) Representation of the pictures and tables in papers in the field of geoscience. Concept *papertable* has **3** data properties including table's tag, caption and dpi.

Knowledge (ace:knowledge) Representation of the item that can express inherent key information in papers in geoscience. Concept *ace:knowledge* has **3** data properties include the name (reusing axiom *skos:prefLabel*), definition (reusing axiom *skos:definition*) and original source. We will introduce the knowledge extraction method amply in the next section.

Location (geo:location) Representation of the geographical, social, political locations. Concept *geo:location* has **3** data properties including location name (reusing axiom *rdfs:label*), latitude and longitude.

Timescale (geo:timescale) Representation of the geologic time scale. Concept *geo:timescale* has **1** data properties including timescale name (reusing axiom *rdfs:label*).

Geohash (http://geohash.org/) Representation of the GeoHash value for the coordinate location. We generate a hash value of the coordinate based on the Geohash algorithm and reuse it as a geohash concept. Based on this hash value, we locate the geographic location of the research articles.

Relations. Also can be deemed for concepts' object properties. The axioms corresponding to relations are defined as following:

acer:is_cited_by connects two paper concepts. It means that the latter paper refers to the former one.

acer:on_the_topic_of connects concept paper and concept topic, and it shows what topic the paper is about.

acer:is_written_by connects concept paper and concept author, which illustrate that who wrote the paper.

acer:is_published_in connects concept paper and concept journal shows that the journal in which the paper is published.

acer:has_illustration and acer:has_table connect concept paper and concept illustration or papertable, which means that the paper contains the picture or the table.

acer:mention_location connects concept paper and concept location, showing that the paper mentions the location.

acer:mention_timescale connects concept paper and concept timescale, showing that the paper is talking about the phenomenon at that timescale.

acer:is_last_known_in connects concept author and concept affiliation, showing that the author works in the affiliation when he/she published his/her last paper.

acer:is_located_in connects concept affiliation and concept location, which means that a scientific research affiliation is located in a geographic location.

acer:has_geohash not only connects concept illustration and geohash entity but also connects concept location and geohash entity. In this way, we shed light on the relationship between geoscience papers and geographical maps.

geor:in_the_period_of and geor:before connect two concept timescale. *geor:in_the_period_of* means that one era happens during the period of the other, while *geor:before* claims that one era happens before the other.

#sameAs , driven by the Semantic Web LinkedData Project, connects concept topic, knowledge, affiliation, and country in GAKG to Wikidata, thereby expanding the scope of usage of GAKG, and also facilitating it to be retrieved and queried in conjunction with other KGs.

acer:mention_knowledge is upper class object properties of relations acer:has_concluded, acer:has_theme, acer:has_designed, acer:learn_in_the_way_of and acer:has_developed, connects concept paper and concept knowledge. It is used to express papers' conclusions, themes, design ideas, and learning methodologies as well as the progress of the disciplines promoted by papers. This part will be introduced in detail in Section 3.

2.2 Statistics of GAKG Datasets

GAKG dataset is preserved in the format of RDF (N-Triple) and currently consists of **68,629,515** triples, including **8,991,737** concepts instance and **41,664,304** links. We provide **271,156** *#sameAs* axioms linking paper's topics and affiliations to the entities in WikiData. The statistics of all the GAKG's entities and links are shown in Table 1 while relations are in Table 2.

Among the statistics above, four social networks are extracted. First, the collection of relation *is_cited_by* is a citation network in the field of geoscience. Second, the collection of *on_the_topic_of* is a bipartite network of papers and topics. Third, the collection of *is_written_by* is a bipartite network of papers and authors, and it can be converted to an author's cooperation network. Figure 3 and Table 3 show the relevant attributes of these four social networks. Moreover, we build the citation network and the cooperation network as community detection benchmarks, which would be detailedly stated in Section 3.

Table 1: Statistics of GAKG Concepts (Up to May 30, 2021).

Concept	Count	Concept	Count
paper	1,122,094	knowledge	62,576
author	908,933	illustration	3,562,816
affiliation	27,175	papertable	760,054
topic	765,184	location	784,279
journal	194	geohash	996,731
timescale	1,701	**Total**	8,991,737

Table 2: Statistics of GAKG Relations (Up to May 30, 2021).

Relation	Count	Relation	Count
is_cited_by	17,704,495	mention_knowledge	704,899
on_the_topic_of	10,401,972	mention_location	759,260
is_written_by	3,547,077	has_geohash	1,021,870
is_published_in	1,122,094	mention_timescale	1,120,398
is_last_known_in	662,850	in_the_period_of	189
is_located_in	25,019	before	155
has_illustration	3,562,816	#sameAs	271,156
has_table	760,054	**Total**	41,664,304

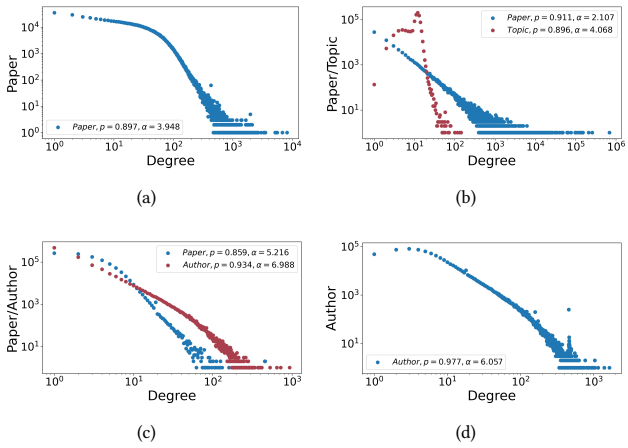


Figure 3: Degree Distribution of the Networks Generated by GAKG, including (a) Citation Network, (b) Paper and topic's Network, (c) Paper and Author's Network and (d) Coauthor Network.

2.3 GAKG SPARQL Endpoint

To help researchers in the field of semantic web and geoscience data mining explore GAKG, we provide a SPARQL Endpoint built upon a Virtuoso triplestore on https://www.acekg.cn/sparql, a simple endpoint for SPARQL query. For convenience sake, we adopt SNORQL to construct a user-friendly SPARQL endpoint on https://snorql.acemap.cn/, so that related scholars can view the link to each entity of each type of ontology including the pictures, tables, and relevant geographic information of articles. The SPARQL Query prefixes and a query example are introduced in Figure 4(a) is the

Table 3: Networks Generated by GAKG.

Network/Relation	Concept	Size	Volume	Max Degree	Avg. Degree	α	p	x_{min} [1]
Co-author Network	author	752,718	5,231,507	1,648	13.900	6.057	0.977	500
Citation Network	paper	884,421	17,704,495	8,165	38.930	3.948	0.897	1,237
is_written_by	paper	1,027,153	3,410,468	458	3.320	5.216	0.859	63
	author	908,902		932	3.752	6.988	0.934	238
on_the_topic_of	paper	944,052	10,182,977	144	10.786	4.068	0.896	33
	topic	89,154		690,961	114.217	2.107	0.911	16,430

[1] α, p, x_{min} is used to evaluate whether the degree distribution of the above four networks is a powerlaw distribution.

Figure 4: Query over GAKG SPARQL Endpoint.

PREFIX list that can be used in our SPARQL endpoint, 4(b) is a sample query example and 4(c) is the result of 4(b). Besides, we also provide a function of generating SPARQL queries according to the keywords input by users on SPARQL Endpoint, where contains more details and examples.

3 BUILDING MULTIMODAL GAKG

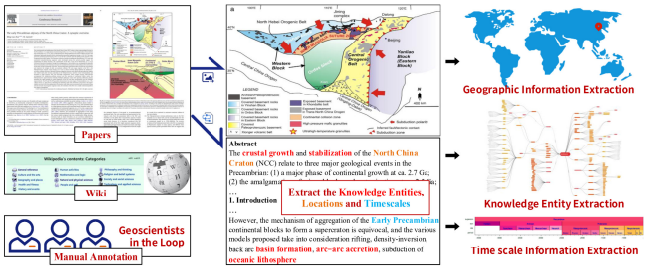


Figure 5: Multimodal Information Extraction for GAKG.

For the management of every piece of knowledge and information mentioned in papers in the field of geoscience, we fuse every single papers' illustrations, tables, and their mentioned knowledge entities, timescale as well as geographic locations to construct a multimodal Knowledge Graph. In this section, we introduce the pipeline of knowledge entities extraction, knowledge taxonomy building, geologic time scale information mining, and geographic information extraction. The pipeline of these processes is shown in Figure 5.

3.1 Knowledge Entity Extraction

In this paper, we propose a Human-In-The-Loop knowledge entity extraction method. First of all, we have to clarify what core knowledge fragments can be abstracted as Knowledge Entities.

Definition 3.1. **Knowledge Entity**: A knowledge entity is a conceptual entity that can highly summarize a group of words with similar meanings, with an unambiguous label and definition. In a sentence, the knowledge entity can express the main idea of the sentence to some extent.

Refer to the division of geographic issues by IGU [4], combined with the idea of a one-sentence summary of scientific research papers from AceMap, we divide the paper's internal knowledge points in the field of geoscience into 5 categories, which are also five questions that geoscience researchers need to answer when understanding a research paper in their field. The five-question is stated as follows and statistics of the subclass of relation *acer:mention_knowledge* are shown in table 4.

- **What is the conclusion of this paper?** Corresponding to the relation *has_concluded*, an object property of paper.
- **What is the theme of the paper?** Corresponding to the relation *has_theme*, an object property of paper.
- **What has the paper designed?** Corresponding to the relation *has_designed*, an object property of paper.
- **What is the research method of the paper?** Corresponding to the relation *learn_in_the_way_of*, an object property of paper.
- **What development in the field of geoscience has the paper promoted?** Corresponding to the information of the relation *has_developed*, an object property of paper.

Table 4: Statistics of *mention_knowledge* Subclass Relations.

Relation Types	Count
has_concluded	203,860
has_designed	6,907
has_developed	45,132
has_theme	283,117
learn_in_the_way_of	165,883

In this way, we define 5 kinds of relations between concept knowledge and concept paper. The pipeline for knowledge entity extraction is introduced as follows.

First, we answer the above five questions from the papers' abstract. Based on the idea of ERNIE [43], we calculate the embedding of each paper entity by deploying network embedding on GAKG's citation network, fuse them with the embedding vectors of the annotated article abstracts by using a pre-trained language model BERT, then use ERNIE's framework and the annotation of the answers to these five questions on 2000 papers' abstracts by geoscience experts to train a machine reading comprehension model and finally generate machine reading answers for the rest of the articles. In this way, for each paper, we get questions and answers pairs.

Second, we associate the answers of machine reading with the knowledge entities. We aggregated 2,377,059 concepts from AceMap

and DBpedia. The definition of each entity is different from the others. What we want is to extract the entities mentioned in the answers. Referring to explicit semantic analysis [8], where a word is represented as a column vector in the TF–IDF matrix of the text corpus and a document is represented as the centroid of the vectors representing its words, we take 2.4 Million concept entities and their descriptions as documents, denoted as D, and take the answers to five questions for each article as queries (each paper has up to five questions). So, for each pair of question and paper, denoted as q, we have obtained a number of candidate entities $E = \{e_i\},\ i \geq 0$. This step can be summarized as equation 1.

$$E = Query(q,\ D), \quad E = \{e_i\},\ i \geq 0 \tag{1}$$

Finally, we rank the candidate entities for each q to obtain the top-3 most similar entities to the answers to five questions. We first construct the features of these candidate entities by calculating the similarity score between the entities with title, abstract of the original papers, and the answers after machine reading based on their TF-IDF scores, and the length, complexity, and letters' amount of the entities. We combine the 6 features above as entity feature vectors. Then, referring to learning to rank algorithms, LambdaRank [26], we try to learn the function shown by equation 2.

$$f(q,\ E) = S, \quad E = e_i, S = s_i,\ i \geq 0 \tag{2}$$

which means that given a paper and a question pair q and a list of candidate entities $E = \{e_i\},\ i \geq 0$, we can generate a score set $S = \{s_i\},\ i \geq 0$, which is used to rank. Therefore, we train a 2-layer neural network model with entities features and a loss function as equation 3:

$$Loss_{ij} = \log\left\{1 + exp\left(-\sigma\left(s_i - s_j\right)\right)\right\} \cdot \left|\Delta NDCG_{ij}\right|, \tag{3}$$

where σ is a parameter determines the shape of the sigmoid function, s_i and s_j are a pair of score we predict and $\Delta NDCG_{ij}$ denotes the difference when exchanging the order of e_i and e_j, where $NDCG$ [11] is an evaluation metric of sorting. $NDCG$ score is used to evaluate the accuracy of ranking, so that its value needs to be as large as possible. The label we used in calculating $NDCG$ is marked by geoscience experts according to every pair of paper and its questions pair q and candidate entities e generated by ESA step.

After training, to ensure that the mined entities can precisely correspond to the knowledge points in the paper, we set a threshold to ensure the model has high precision. Based on this threshold, the precision of our model on the benchmark we set is 0.92, and the recall rate is 0.34. The pipeline of the ranking model is shown in Figure 6.

3.2 Geoscience Knowledge Taxonomy

Scientific articles are rich in the knowledge entities of disciplines, and these knowledge entities can be connected by hyponymy relations. Geoscience articles are no exception where existing plentiful knowledge entities, as well as subordination and inheritance relations between entities. Since geoscience accumulates a huge knowledge system and numerous academic papers, therefore, building a knowledge taxonomy for geoscience is essential. With the help of senior geoscientists, we combine the taxonomy of the category in Wikipedia and relations between the academic fields from AceMap to construct the hierarchical structure of geoscience knowledge

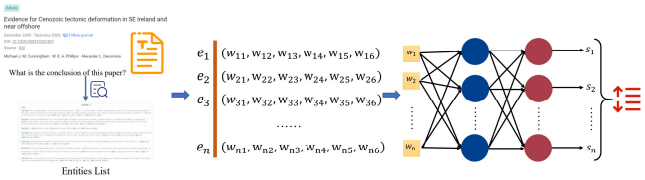


Figure 6: The Workflow of the Ranking Model.

entities and map the knowledge entities extracted by our Human-In-the-Loop system to this taxonomy. Finally, the geoscience knowledge taxonomy is established.

3.3 Geographical Information Extraction

In the areas of geoscience, scientific articles' geographical information is abundant. For one thing, because of the discipline characteristics, a large number of geoscience papers have geographic illustrations, where geographical coordinates are shown so that the places the papers mentioned can be seized. For another, distinguish from other subjects, the locations would be directly shown in the text, indicating the research locations or reference locations of the articles. Consequently, we make effort to extract geographic illustrations as well as locations from the text and get their location coordinates.

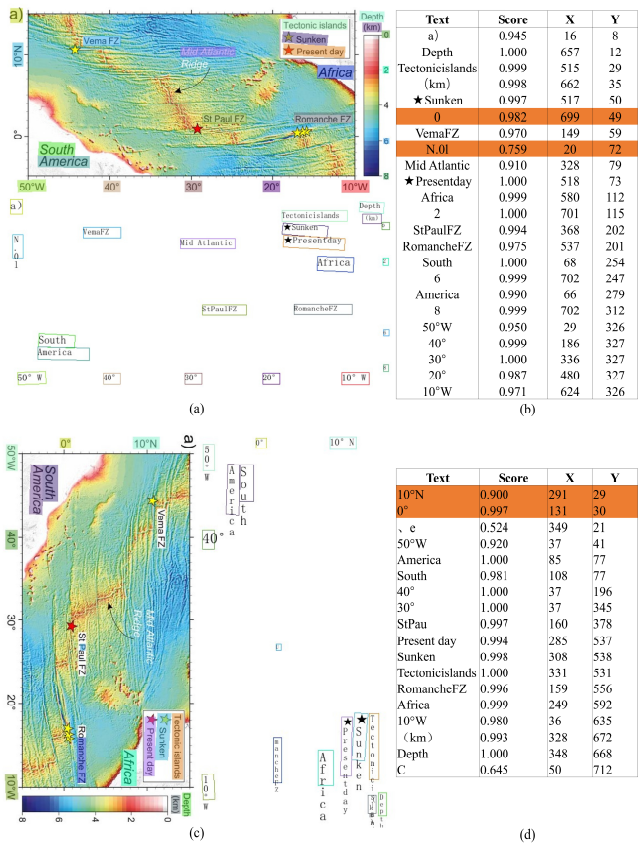
Geographic illustrations. We first use pdffigures2 [5] to extract pictures from PDFs from papers with open access certification, perform rule-based screening on the extracted pictures according to the picture features. Second, we not only extract the text representing place names from these illustrations, and generate the coordinates with geocoder pipeline, but also extract the latitude and longitude range from the numbers representing the latitude and longitude from these illustrations. Throughout this process, we perform image enhancement and flipping in order to have a better image recognition performance with PaddleOCR. Moreover, we annotate 1000 illustrations as the ground truth benchmark and adjust the input of the PaddleOCR by using rule-based methods (illustrated in Figure 7) to ensure the accuracy on our benchmark to reach above 85%. Finally, we apply the method to all the pictures in GAKG to generate a geographic coordinate.

Geographical/Social/Political Entities. We build a BERT-based named entity recognition model to extract the locations and GPE (Geographical/Social/Political Entities). After the normalization of the location entities, we use a geocoder pipeline to get the location coordinates.

Finally, we use the Geohash algorithm to store coordinates for the sake of the visualization of SPARQL query results.

3.4 Geologic Time Scale Extraction

In addition to the extensive distribution in geographic space, the research of geosciences also has historical continuity. To more conveniently use the research findings of geoscientists to discover the information of the geographic era involved, we extract the geologic time scale entities mentioned in the title, abstract and introduction of the papers through a rule-based enhance information extraction method that we developed. According to the geographic era words' positions in papers and the words appearing before and after, the confidence scores are calculated. Finally, we map the geologic



Text	Score	X	Y
a)	0.945	16	8
Depth	1.000	657	12
Tectonicislands	0.999	515	29
(km)	0.998	662	35
★Sunken	0.997	517	50
0	0.982	699	49
VemaFZ	0.970	149	59
N.0l	0.759	20	72
Mid Atlantic	0.910	328	79
★Presentday	1.000	518	73
Africa	0.999	580	112
2	1.000	701	115
StPaulFZ	0.994	368	202
RomancheFZ	0.975	537	201
South	1.000	68	254
6	0.999	702	247
America	0.990	66	279
8	0.999	702	312
50°W	0.950	29	326
40°	0.999	186	327
30°	1.000	336	327
20°	0.987	480	327
10°W	0.971	624	326

Text	Score	X	Y
10°N	0.900	291	29
0°	0.997	131	30
、e	0.524	349	21
50°W	0.920	37	41
America	1.000	85	77
South	0.981	108	77
40°	1.000	37	196
30°	1.000	37	345
StPau	0.997	160	378
Present day	0.994	285	537
Sunken	0.998	308	538
Tectonicislands	1.000	331	531
RomancheFZ	0.996	159	556
Africa	0.999	249	592
10°W	0.980	36	635
（km）	0.993	328	672
Depth	1.000	348	668
C	0.645	50	712

Figure 7: An Example of Illustrations Coordinates Extraction. (a) is the result of OCR, (b) highlights the recognition error, (c) is the result of OCR after using a rule-based method adjusting image, and (d) highlights the corrected coordinates.

time scale words that reach a certain threshold to the normalized high-precision geologic time scale vocabulary collected by senior geoscientists. We have extracted more than 250k articles that have geographic time scale information. Based on the evolution time of the age, the most fine-grained time scales, the distribution of these papers is shown in Figure 8.

4 BENCHMARKS

In this section, we discuss two kinds of benchmarks including three datasets that we provide with the GAKG, including a citation network and a cooperation network for community detection task as well as a tiny KG extracted from GAKG for knowledge representation learning task. We discuss the details of these benchmark datasets as follows. All the benchmarks can be downloaded from the GAKG Github repository[2].

4.1 Community Detection

Community detection is one of the classic issues in social network analysis, targeting to find nodes' groups, in some sense, one node has more connection within its group than the others. There

[2] https://github.com/davendw49/gakg

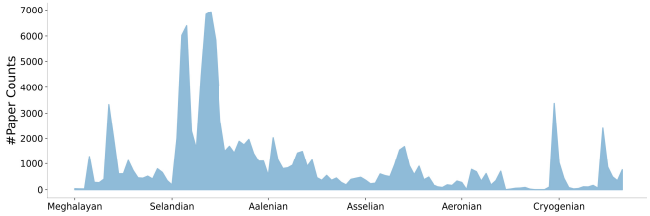


Figure 8: Papers Distribution along with Geologic Era.

are plenty of works to find community in undirected networks [6, 15, 23, 27], while few works in directed networks [24].

Community detection over the undirected network is a classic network science problem. A large number of works can successfully perform well on specific benchmarks. However, most of the existing benchmarks remain for model evaluation and have little practical significance. Community detection on social networks in the real world needs more sociological meanings benchmarks. When it comes to clustering problems in directed networks, the issue has not received attention from the scientific community. [19] The directed network datasets with ground-truth communities are rare, like email-Eu-core [16] and cora [41]. DBLP [34] provides a large-scale citation network among all the fields however it lacks community labels, which need to be generated by researchers.

4.1.1 Datasets. **GeoScience Papers Citation Network (GPCN).** Based on GAKG, we construct a GeoScience Papers Citation Network (GPCN), where papers are denoted as nodes while edges' direction means one paper refers to the other paper. GPCN is a weakly connected directed graph, with 1,598 weakly connected components, and the largest weakly connected component has 838,219 nodes and 16,031,892 edges (both nodes and edges coverage are over 0.99). Besides, each paper has a community label according to the journal in which it was published.

GeoScience Authors Cooperation Network (GACN). Based on GAKG, we also build a GeoScience Authors Cooperation Network (GACN), where authors are denoted as nodes while edges mean two authors have cooperation relation. Each edge has a weight denoting two authors' cooperation times. GACN is a unconnected graph, with 32,863 connected components, and the largest one has 752,718 nodes and 5,231,507 edges (containing 87.5% nodes and 97.2% edges). Besides, each author has a disjoint community label refer to the journal the author has published most articles on it.

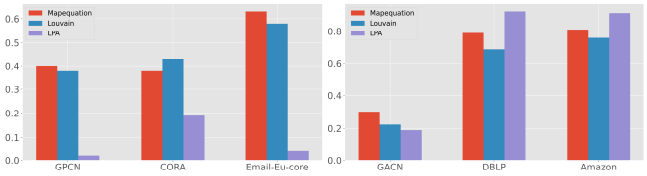
GPCN and GACN can be accessed via GAKG resources pages, and their statistics are shown with other benchmarks in Table 5.

4.1.2 Experiments and Results. For community detection problems on both the directed and undirect network, we compare Mapequation [29], Louvain[2], and LPA [28] over 3 directed network, Email-Eu-core [16], CORA [20], and GPCN, and 3 undirected network, DBLP Collaboration Network, Amazon Product Network [40], and GACN. For the record, we use the GPCN's largest weakly connected component and the GACN's largest connected component as the input data when evaluating models over GPCN and GACN.

We observe two kinds of community detection tasks using NMI [31] as the evaluation metric and the results are shown in Figure 9. According to the results, it is clear that these algorithms do not perform well on our benchmarks, even though the algorithms

Table 5: Statistics of Community Detection Benchmarks.

Benchmarks	Number of communities	Nodes	Edges	Nodes in Largest WCC	Edges in Largest WCC	Nodes in Largest SCC	Edges in Largest SCC	Average Cluster Coefficient	Triangles	Diameter
GPCN	194	842,121	16,034,510	838,219 (0.995)	16,031,892 (0.999)	0	0	0.0699	38,789,469	176
GACN	194	860,280	5,381,861	752,718 (0.875)	5,282,032 (0.972)	752,718 (0.875)	5,282,032 (0.972)	0.6897	43,502,542	15
Email-Eu-core	42	1,005	25,571	986 (0.981)	25,552 (0.999)	803 (0.799)	24,729 (0.967)	0.3994	105,461	7
CORA	7	2,708	5,429	2,485 (0.918)	2,604 (0.493)	13 (0.005)	14 (0.003)	0.1314	1,630	15
DBLP (Collaboration Network)	2,547	317,080	1,049,866	317,080 (1.000)	1,049,866 (1.000)	317,080 (1.000)	1,049,866 (1.000)	0.6324	2,224,385	21
Amazon (Product Network)	5,000	334,863	925,872	334,863 (1.000)	925,872 (1.000)	334,863 (1.000)	925,872 (1.000)	0.3967	667,129	44



(1) Directed Network Community Detection (2) Undirected Network Community Detection

Figure 9: Community Detection Evaluation Results.

are efficient and suitable for detecting communities over large-scale networks. Therefore, in the task of community detection, our benchmark **GPCN** and **GACN** put forward a higher challenge. Moreover, it is also very significant to use GPCN and GACN to discover research groups in the world of geoscience research.

4.2 Knowledge Representation Learning

Knowledge representation learning (KRL) has always been a hot research area, and many extraordinary works [3, 14, 25, 32, 39] have performed well on link prediction task over FB15K and WN18 [3, 21] benchmark.

4.2.1 Dataset. Even though knowledge representation models can be evaluated over FB15K and WN18, the models struggle with academic KG, since academic KG representation learning lacks appropriate benchmarks. Therefore, we extract a dataset from GAKG, named **GA16K** and evaluate the current SOTA algorithm on it. We first order each type of entity according to their degrees and select the entities with a larger degree into an entity set V. Then we add edges into relation set R if its source node and target node are in the set V. And the train, test, and valid data sets are randomly divided. Table 6 shows the basic statistics of FB15K, WN18 and GA16K.

4.2.2 Experiments and Results. We use the link prediction task defined in TransE [3] for evaluation. Assuming that a triple (h, r, t) in a KG consists of two entities, $h, t \in V$ and relation $r \in R$, the algorithm embeds the entities in the k-dimensional space and based on the prediction of t given h and r, or the result of h given r and t, calculate *MR*, the average rank of correct entities, and *hit@10*, the proportion of correct entities appearing in the top 10, to evaluate the performance of the algorithms. We compare the performance of these algorithms over the three benchmarks (FB15K, WN18 and GA16K). The algorithms we take into evaluation are RESCAL [25], TransE [3], TransH [39], SimplE [14] and RotatE [32].

The evaluation results on GA16K are produced with optimal training parameters of each model, and the results on FB15K and WN18 are extracted from the models' original papers. The results are shown in Table 7. We can see that GA16K has the same trend as FB15K and WN18. Among the translation-based models, we compared, from TransE, TransH to RotatE, it is getting better and better, indicating that translation-based is effective on GA16K. The performance of SimplE on GA16K is not as great as it is on other

Table 6: Statistics of KRL Benchmarks.

Benchmark	relation	entity	triple
FB15K	1,345	14,951	483,142
WN18	18	40,943	141,442
GA16K	10	16,363	151,662

Table 7: Results of Link Prediction Task.

Models	FB15K		WN18		**GA16K**	
	MR	hit@10	MR	hit@10	MR	hit@10
RESCAL	683	0.441	1,163	0.528	4,300	0.001
TransE	125	0.471	251	0.892	280	0.320
TransH	84	0.585	303	0.867	337	0.325
RotatE	40	0.884	309	0.959	214	0.366
SimplE	74	0.876	412	0.947	311	0.260

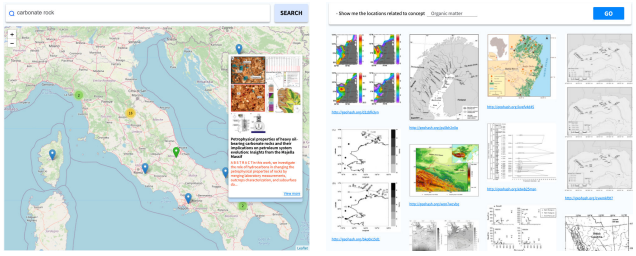
benchmarks to some extent. The reason is that SimplE aims at the canonical polyadic decomposition of entities in triples, but it is not suitable for GA16K, since that if exchange the head entities and tail entities, only one kind of edge, *is_cited_by*, will not be the wrong edge, the rest are all wrong, which is a kind of noise in the training process.

Comprehensive speaking, all the algorithms perform not quite well on GA16K. GA16K is based on uniformly sample node and all the relevant edges, acting as a subgraph of GAKG, GA16K is more challenging on knowledge representation learning tasks than other benchmarks. Compare to other benchmarks, except for the paper entities, other entities only have one out-edge and one in-edge, so the embedding of the paper entities are influential on the whole training process since the paper entities are always the head entities and only become tail entities when they are connected by *is_cited_by*. Therefore, in the task of link prediction, our benchmark GA16K puts a new challenge on the current knowledge embedding models.

5 ONLINE APPLICATIONS

Benefit from the multimodality of GAKG, computer scientists investigate more about information retrieval and data mining techniques on geoscience, while geoscientists conduct research in a more visual way. As a resource for the information technology community, we provide two examples of applications based on the GAKG. The screenshots of the demos are shown in Figure 10.

Geographic Information Retrieval. We provide a knowledge-based search engine on a geographical map for the literature of geoscience. First, researchers can query the paper-oriented information in GAKG, including papers' titles, abstracts, timescale, topics, and knowledge entities. Researchers enter keywords in the input box, *e.g.* "Carbonate rock", and the relevant papers would be shown on the map (a demo screenshot is shown in 10(a)). Once there are plentiful articles in the text that contain the input keywords, or



(a) Geographic Information Retrieval (b) Geoscience KBQA

Figure 10: GAKG Application Demos.

there are many articles' topics or knowledge entities that are the input keywords, a great number of papers will be returned. For a better user experience, we only show the distribution of papers in the display area of the window. If the researcher drags the window, the distribution of the papers will change accordingly.

Geoscience KBQA. Based on GAKG, research can know more information about the relation between papers. We carry out several samples including one-hop queries, such as returning papers targeting a particular topic, two-hop queries, such as querying illustrations in a specific field, three-hop queries, such as querying geographic locations that a certain affiliation often studies and querying the relationship between geographic locations and affiliations is a kind of four-hop queries. These template-based queries can be applied in scientific research and academic communication. These questions are also generally inextricable by existing Q&A systems and search engines. (a demo screenshot is shown in 10(b))

6 DISCUSSIONS

In this section, we discuss the value of data, issues, and other observations regarding the construction and applications of GAKG.

6.1 Availablility and Quality

Availablility. The dump files of GAKG, stored in N-Triples format, are available at the GAKG homepage: https://gakg.acemap.info /. GAKG can be queried and explored via the SPARQL endpoint. The implementation of the pipeline to extract knowledge entities of GAKG, the benchmarks and the source code of the baselines' implementation are available at the GAKG Github repository.

Quality. During constructing the GAKG, We have been ensuring the quality of the KG included by manual verification of some core data and extracting data from reliable sources. Rigorously, we merge AceMap, DBpedia, and DDE's vocabulary to build an entity library, checking periodically by experts in the domain of geoscience. In future work, we will continue to improve our entity library's quality to make GAKG more precise, serving data mining tasks such as relation extraction and reasoning.

6.2 Limitations

Despite the high coverage and value in the use of GAKG, its precision is limited by knowledge entity extraction. Unsupervised information extraction systems do not have a good performance on scientific text, because of the lacking annotation about the knowledge entities and their relationships within the papers. [10]. Therefore, the recall of our model was reduced to ensure the high precision of the extracted entities. Apart from that, since we keep human experts' annotation in the loop, the amount of annotations needs to be accumulated so that the model's performance can be improved by iterations.

6.3 Future Directions

We will maintain and update the GAKG in the future with the promotion of GAKG and its query system sustainedly. Based on GAKG, the observation directions in the future we expect and follow are listed as follows:

Academic Knowledge System Construction with Human-In-the-Loop. We propose a Human-In-the-Loop framework in this paper to absorb the knowledge entities of each paper transferring the unstructured data into graph data. It stands to reason that mining the connection of two knowledge entities is much more important. Therefore, we intend to mine the knowledge and construct the knowledge systems of disciplines based on academic articles' knowledge entities and their relations, with Human experts' annotation in the loop of academic knowledge system construction.

Social Community Detection in Geoscience. Despite sharing community detection benchmarks to evaluate the performance of models and algorithms, finding the collaboration between geoscientists and seeking influential papers among the citation networks rather than only considering their citation.

Scientific Articles' Geographical Information Extraction. In this paper, we have collected the locations mentioned by papers, while the geographic positions the papers study may be omitted. Therefore, it is important to sort out the location the geoscience papers investigate, which sheds light on the connection between scientific articles and geographical maps.

7 CONCLUSIONS

In this work, we present a novel multimodal academic knowledge graph for geoscience, named GAKG, to facilitate the geoscience experts to find the internal relations between the articles, the geographic location, and every scrap of the knowledge in papers. GAKG brings out a pipeline to find key knowledge entities from the scientific articles and connects these entities to the paper entities with geoscientists in the loop. In addition, GAKG extracts the geographic location information and geologic time scale entities to expand the dimension of papers parsing papers' illustrations and text. Based on the multimodal data, we build a SPARQL endpoint and a search engine for research to explore GAKG. We also provide benchmarks for community detection and knowledge representation learning tasks. Finally, we discuss the applications, availability, quality, limitations, and future directions of GAKG.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China (No.2018YFB2100302), NSF China (No.42050105, 61960206002, 61822206, 62020106005, 61829201), 2021 Tencent AI Lab RhinoBird Focused Research Program (No: JR202132) and Shanghai Academic/Technology Research Leader Program (No. 18XD1401800).

REFERENCES

[1] Surendra Adhikari, Lambert Caron, Bernhard Steinberger, John T Reager, Kristian K Kjeldsen, Ben Marzeion, Eric Larour, and Erik R Ivins. 2018. What drives 20th century polar motion? *Earth and Planetary Science Letters* 502 (2018), 126–132.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*. 1–9.

[4] IGU CGE. 1992. International charter on geographical education. *International Geographical Union, Commission on Geographical Education* (1992).

[5] Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. IEEE, 143–152.

[6] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

[7] National Research Council et al. 2001. *Basic research opportunities in earth science*. national academies Press.

[8] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)* 29, 2 (2011), 1–34.

[9] Jun-xuan Fan, Shu-zhong Shen, Douglas H Erwin, Peter M Sadler, Norman MacLeod, Qiu-ming Cheng, Xu-dong Hou, Jiao Yang, Xiang-dong Wang, Yue Wang, et al. 2020. A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. *Science* 367, 6475 (2020), 272–277.

[10] Paul Groth, Mike Lauruhn, Antony Scerri, and Ron Daniel Jr. 2018. Open Information Extraction on Scientific Text: An Evaluation. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3414–3423. https://www.aclweb.org/anthology/C18-1289

[11] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[12] Amar Viswanathan Kannan, Dmitriy Fradkin, Ioannis Akrotirianakis, Tugba Kulahcioglu, Arquimedes Canedo, Aditi Roy, Shih-Yuan Yu, Malawade Arnav, and Mohammad Abdullah Al Faruque. 2020. *Multimodal Knowledge Graph for Deep Learning Papers and Code*. Association for Computing Machinery, New York, NY, USA, 3417–3420. https://doi.org/10.1145/3340531.3417439

[13] Yasuhiro Kato, Koichiro Fujinaga, Kentaro Nakamura, Yutaro Takaya, Kenichi Kitamura, Junichiro Ohta, Ryuichi Toda, Takuya Nakashima, and Hikaru Iwamori. 2011. Deep-sea mud in the Pacific Ocean as a potential resource for rare-earth elements. *Nature Geoscience* 4, 8 (2011), 535–539.

[14] Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. *arXiv preprint arXiv:1802.04868* (2018).

[15] Andrea Lancichinetti and Santo Fortunato. 2009. Community detection algorithms: a comparative analysis. *Physical review E* 80, 5 (2009), 056117.

[16] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 2–es.

[17] Jiaqi Liu, Luoyi Fu, Yuhang Yao, Xinzhe Fu, Xinbing Wang, and Guihai Chen. 2018. Modeling, analysis and validation of evolving networks with hybrid interactions. *IEEE/ACM Transactions on Networking* 27, 1 (2018), 126–142.

[18] Jiaqi Liu, Qin Zhang, Luoyi Fu, Xinbing Wang, and Songwu Lu. 2019. Evolving knowledge graphs. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2260–2268.

[19] Fragkiskos D Malliaros and Michalis Vazirgiannis. 2013. Clustering and community detection in directed networks: A survey. *Physics reports* 533, 4 (2013), 95–142.

[20] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval* 3, 2 (2000), 127–163.

[21] George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

[22] Ming-Guo Zhai and M. Santosh. 2011. The early Precambrian odyssey of the North China Craton: A synoptic overview. *GONDWANA RESEARCH* 1, 6–25. https://deep.acemap.info/paper/366752

[23] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113.

[24] Mark EJ Newman and Elizabeth A Leicht. 2007. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences* 104, 23 (2007), 9564–9569.

[25] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Icml*.

[26] C Quoc and Viet Le. 2007. Learning to rank with nonsmooth cost functions. *Proceedings of the Advances in Neural Information Processing Systems* 19 (2007), 193–200.

[27] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. 2004. Defining and identifying communities in networks. *Proceedings of the national academy of sciences* 101, 9 (2004), 2658–2663.

[28] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 76, 3 (2007), 036106.

[29] Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4 (2008), 1118–1123.

[30] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*. 243–246.

[31] Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* 3, Dec (2002), 583–617.

[32] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197* (2019).

[33] Svetlana L. Stuefer, Christopher D. Arp, Douglas L Kane, and Anna K. Liljedahl. 2017. Recent Extreme Runoff Observations From Coastal Arctic Watersheds in Alaska. *WATER RESOURCES RESEARCH* 11, 9145–9163. https://deep.acemap.info/paper/738179

[34] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 990–998.

[35] Timothy M. Kusky, Brian F. Windley, Lu Wang, Zhensheng Wang, Xiaoyong Li, and Peimin Zhu. 2014. Flat slab subduction, trench suction, and craton destruction: Comparison of the North China, Wyoming, and Brazilian cratons. *Tectonophysics* 630, 208–221. https://www.acemap.info/paper/233963873

[36] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, et al. 2020. COVID-19 literature knowledge graph construction and drug repurposing report generation. *arXiv preprint arXiv:2007.00576* (2020).

[37] Ruijie Wang, Yuchen Yan, Jialu Wang, Yuting Jia, Ye Zhang, Weinan Zhang, and Xinbing Wang. 2018. Acekg: A large-scale knowledge graph for academic data mining. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1487–1490.

[38] Shu Wang, Xueying Zhang, Peng Ye, Mi Du, Yanxu Lu, and Haonan Xue. 2019. Geographic knowledge graph (GeoKG): A formalized geographic knowledge representation. *ISPRS International Journal of Geo-Information* 8, 4 (2019), 184.

[39] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14)*. AAAI Press, 1112–1119.

[40] Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42, 1 (2015), 181–213.

[41] Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. 2017. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 555–564.

[42] Yi Zhang, Yong Gao, LuLu Xue, Si Shen, and KaiChen Chen. 2008. A common sense geographic knowledge base for GIR. *Science in China Series E: Technological Sciences* 51, 1 (2008), 26–37.

[43] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129* (2019).

[44] Chenghu Zhou, Hua Wang, Chengshan Wang, Zengqian Hou, Zhiming Zheng, Shuzhong Shen, Qiuming Cheng, Zhiqiang Feng, Xinbing Wang, Hairong Lv, et al. 2021. Prospects for the research on geoscience knowledge graph in the Big Data Era. *Science China Earth Sciences* (2021), 1–11.