International Journal of
Geo-Information

Article

A Knowledge-Driven Geospatially Enabled Framework for Geological Big Data

Liang Wu [1,2], Lei Xue [1], Chaoling Li [3], Xia Lv [3], Zhanlong Chen [1], Baode Jiang [2], Mingqiang Guo [2] and Zhong Xie [1,2,*]

[1] School of Information Engineering, China University of Geosciences, Wuhan 430074, China; wuliang@cug.edu.cn (L.W.); xueleicug@163.com (L.X.); chenzhanlong2005@126.com (Z.C.)
[2] National Engineering Research Center for GIS, Wuhan 430074, China; jiangbaode@mapgis.com (B.J.); gmqandjxs@163.com (M.G.)
[3] Development and Research Center, China Geological Survey, Beijing 100037, China; lchaoling@mail.cgs.gov.cn (C.L.); lxia@mail.cgs.gov.cn (X.L.)
* Correspondence: xiezhong@cug.edu.cn; Tel.: +86-027-6788-3728

Academic Editors: Ozgun Akcay and Wolfgang Kainz
Received: 2 April 2017; Accepted: 31 May 2017; Published: 3 June 2017

Abstract: Geologic survey procedures accumulate large volumes of structured and unstructured data. Fully exploiting the knowledge and information that are included in geological big data and improving the accessibility of large volumes of data are important endeavors. In this paper, which is based on the architecture of the geological survey information cloud-computing platform (GSICCP) and big-data-related technologies, we split geologic unstructured data into fragments and extract multi-dimensional features via geological domain ontology. These fragments are reorganized into a NoSQL (Not Only SQL) database, and then associations between the fragments are added. A specific class of geological questions was analyzed and transformed into workflow tasks according to the predefined rules and associations between fragments to identify spatial information and unstructured content. We establish a knowledge-driven geologic survey information smart-service platform (GSISSP) based on previous work, and we detail a study case for our research. The study case shows that all the content that has known relationships or semantic associations can be mined with the assistance of multiple ontologies, thereby improving the accuracy and comprehensiveness of geological information discovery.

Keywords: geology; ontology; knowledge discovery; spatial data; big data

1. Introduction

Large volumes of geological reports have been accumulated during geological survey procedures, with each report containing different geological themes, such as rocks, minerals, or hydrology. The contents of these reports are stored in different formats, such as .doc, .pdf, .jpg, .tiff, and spatial data files. In addition, these reports consist of large amounts of structured data and unstructured data. Structured data are typically stored and managed using relational or spatial databases; however, the characteristics of unstructured data render them difficult to manage via virtual applications. Unstructured data include diverse types and fragmented information, which contain more abundant information and have more potential value than structured data. Using a traditional file system to manage these data could increase the inefficiency of answering queries and retrieving statistical information and increase the difficulty of retrieving and mining data. Therefore, considerable research is focused on developing methods of efficiently managing, mining and utilizing these unstructured data, and cloud computing and big-data-related technologies have the potential to resolve issues related to unstructured data.

Technical advancements continue to strengthen the relationship between cloud computing and big data. Cloud computing provides a scalable, cost-efficient solution and on-demand processing service for using big data and can address data-oriented challenges [1–3], fostering a potential solution for the transformation of Big Data's 4 Vs (volume, velocity, veracity and variety) into the 5th V (value) [4]. Researchers have combined cloud computing and big-data technologies in different domains. A framework named CIRUS is a generic and elastic cloud-based platform that enables real-time, ubiquitous big-data analytics [5]. Vera-Baquero et al. introduced a cloud-based architecture that leverages big-data technology to support the performance analyses of any business domain and that operates in a timely manner regardless of the underlying issues that are associated with the operational systems [6]. DiploCloud is an efficient, distributed and scalable Resource Description Framework (RDF) data processing system for distributed and cloud environments and uses a resolutely non-relational storage format [7]. Jizhe Xia et al. utilized cloud computing and volunteer computing technologies and proposed a spatiotemporal performance model that provides more accurate performance evaluations to users from different regions at different times [8]. Roberto Giachetta proposed a geospatial data processing framework to enable the management and processing of spatial and remote sensing data in a distributed environment [9]. Big data is a massive set of data that is challenging to manage with traditional applications and includes huge, complex, and abundant structured, semi-structured, and unstructured data alongside hidden data that are generated and gathered from several fields and resources [10]. However, available data-mining techniques are designed for schema-oriented storage and therefore are not applicable to an unstructured data style [11]. The variety and veracity of big data demand new technologies (e.g., Hadoop, HBase) to clean, store, and organize unstructured data [12,13]. The use of NoSQL (Not only SQL) technology is increasing among internet companies and other enterprises to mine information from such diversiform data [14,15]. NoSQL repositories offer great flexibility and speed in terms of data processing, and the key-value style of querying this type of database enables efficient retrieval [16]. TouchR and RSenter are designed to extract terms from unstructured data sources (specifically, NoSQL databases) and are focused on the document-append style of NoSQL storage [17–19]. MapReduce is widely used to improve the performance of systems for large-scale data analyses and various data-management methods [20,21]. For example, Zhong et al. [22] proposed an "indexing + MapReduce" data architecture for efficient spatial query processing. Hadoop-GIS [23] utilized global partition indexing and implicit parallel spatial query execution on MapReduce to achieve efficient query processing. SpatialHadoop [24] is an efficient MapReduce framework for spatial data queries and operations, which builds a two-level spatial index structure and basic spatial components inside the MapReduce layer.

The increased demand for online spatial-information services poses new challenges to the fields of computer science and geographic information science [25]. Because of the explosion of information and changes in the nature of the services and information demanded by users, information technology professionals are finding it difficult to satisfy the needs of their users. Modern users want to retrieve specific information from within the plethora of available information thus, requiring the conversion of simple document text retrieval to knowledge retrieval [26,27]. Ontologies have been used for information system development as one of the main knowledge representation tools. These tools consist of concepts, hierarchies, arbitrary relationships between concepts, and possibly other axioms [28]. Kuo, C.L. and J.H. Hong proposed a new strategy and framework to process cross-domain geodata at the semantic level [29]. This framework leverages the semantic equivalence of concepts between domains through bridge ontology and facilitates the integrated use of different domain data, which has long been considered superior to Geographic Information Systems (GIS).

Some researchers combined ontologies and semantics with geospatial technologies to improve geospatial data or service discovery. Stock, K. et al. described an information model for a geospatial knowledge infrastructure that uses ontologies to represent these semantic details, including knowledge regarding domain concepts, the scientific elements of the resource, and web services, which can be used to enable more intelligent searches over scientific resources and support new methods to infer and visualize scientific knowledge [30]. Cruz, S.A.B. et al. used semantic descriptions of geospatial data quality requirements in a rule-based form. These rules allow the semantic annotation of geospatial data alongside the conditional planning method [31]. Arctic SDI is a prototype that utilizes the knowledge-based approach and spatial web portal technology to propose a hybrid approach for efficient service discovery from distributed web catalogs and the dynamic Internet. This method proposes a domain knowledge base to model the latent semantic relationships among scientific data and services and an intelligent logic reasoning mechanism for (semi-)automatic service selection and chaining [32]. Jung et al. proposed an ontology-enabled framework for a geospatial problem-solving environment that allows collaborations among web service providers, domain experts, and solution seekers to semantically discover and use geographic information services to solve a target class of geospatial problems [33]. Sensor Metadata Ontology [34] was proposed to achieve a unified semantic description for heterogeneous sensors and to promote accurate and efficient discovery. Yingjie Hu et al. [35] developed an ontology for ArcGIS online data to convert metadata into linked data and calibrate a linear regression model for semantic searches and flexible queries for knowledge discovery. In addition, common technologies are widely used when performing semantic-related research. Jena is a free and open source Java framework for building semantic web and linked data applications [36–38], RDF is a framework for describing the available resources and their relationships on a network [28,36,39–41], SPARQL (Simple Protocol and RDF Query Language) is a graph-based query language for RDF [37,42–45], and SWRL (Semantic Web Rule Language) is utilized to provide rules for semantic networks [33,46–48].

At present, in terms of content retrieval and knowledge discovery, many researchers have explored some methods combined with ontology and semantic-related technology. However, many of the current studies are based on structured spatial data, semi-structured metadata, and descriptive spatial services. For large-scale unstructured content, some scholars have done research on storage and information extraction but do not take into account the relevance and fusion value of unstructured data and structured data, thus resulting in the inability to fully excavate the value of unstructured content. Geological big data represent an application of big-data theories and technologies in the geological domain. At present, geological research has transitioned from qualitative research to quantitative research and from scarce data to massive data. Compared to general big data, geological big data contain both massive unstructured data and abundant geospatial information and temporal information, which are significant in the geological sciences. Thus, it is critical to derive diversified information from massive geological survey data, promote geological content retrieval from keyword-based discovery to knowledge-based discovery, and improve the information accessibility of geological data. This paper attempts to explore the methods of applying massive geological data from the perspective of unstructured data, structured data and spatial service integration, which could effectively fuse heterogeneous information and improve the quality of geological content services. In this paper, we establish a knowledge-driven geological survey information smart-service platform (GSISSP) based on our previously developed geological survey information cloud-computing platform (GSICCP) [49]. NoSQL technology is integrated into GSISSP to organize complex unstructured data, and ontologies are imported to build semantic and knowledgeable associations. Our subsequent efforts concentrate on the following aspects of using geological big data: (1) Introduce geological thematic ontology, geological temporal ontology and toponymy ontology (these terms are short for the geological domain ontology), combine big-data storage and processing technologies, split unstructured geologic survey data into fragments, and extract multi-dimensional information from each fragment to rapidly discover target information from massive content. (2) Build associations between geological fragments using relationships presented in the geological domain ontology and promote the retrieval from keyword-based discovery to knowledge-based discovery to improve the accuracy and comprehensiveness of information discovery. (3) Combine semantic-related technologies and establish a retrieval framework ontology. Our work could help transform a specific class of geological problems into workflows to increase the intelligence of the information discovery process.

The remaining sections of this paper are structured as follows. Section 2 proposes the architecture of the GSISSP and provides a brief introduction of the key technologies. Section 3 describes the method for organizing massive unstructured data from geologic surveys in detail. Section 4 introduces a method of associating one content item with another based on geological domain ontology and proposes a question-oriented retrieval framework. Section 5 illustrates a use case. Section 6 presents our conclusions and discusses prospects for future work.

2. Architecture of the GSISSP

In a previous study, we established a GSICCP, which is a platform that provides multiple types of geological services. The GSICCP is vertically divided into five layers: a data layer, a fabric layer, a resources layer, a discovery and integration layer, and an application and representation layer [49]. In this paper, which is based on the architecture of the GSICCP, we intend to store and discover massive geological structured and unstructured data. Hadoop's [50] ecosystem technologies are imported to extend the resources layer, and semantic technologies are imported to extend the discovery and integration layer. Thus, a GSISSP is established to support information mining and discovery. The data layer and fabric layer are maintained in their original conditions, and a geological content discovery portal is built in the application and representation layer. Detailed descriptions of these layers can be found in the literature [49]. The architecture of the GSISSP is shown in Figure 1.

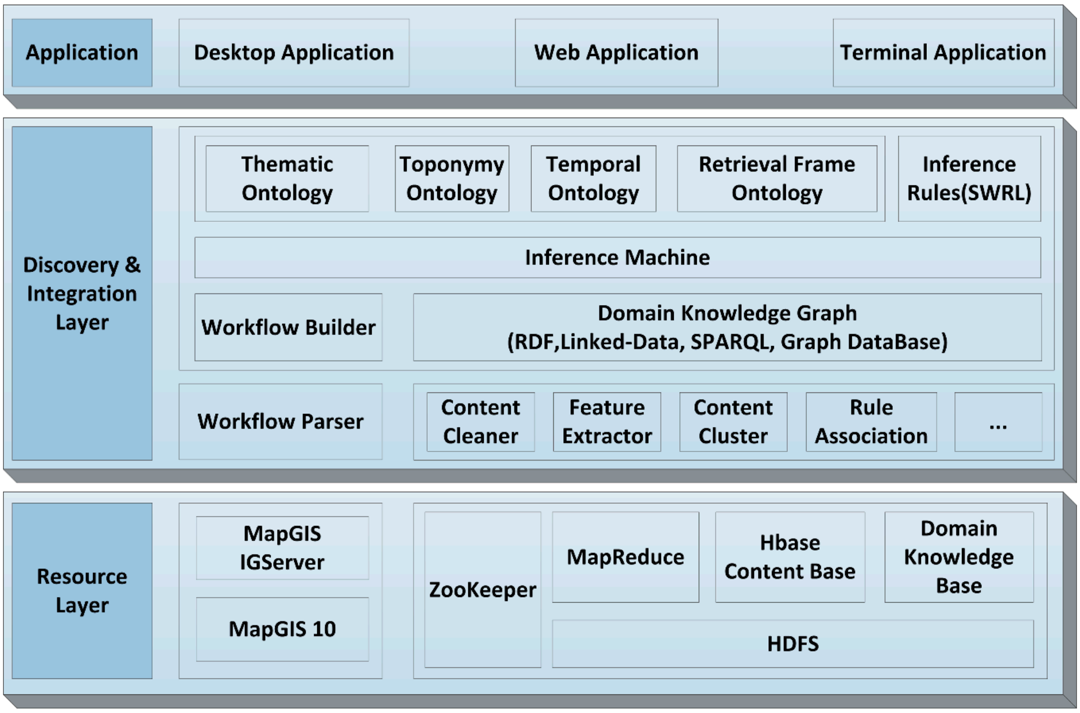


Figure 1. Architecture of the GSISSP.

The resource layer is the foundation of the GSISSP, consisting of certain pivotal modules, including a geological survey information cloud (GSI-CLOUD) resource integration middleware, GSI-CLOUD meta-service libraries, GSI-CLOUD workflow service middleware, and GSI-CLOUD service engine middleware. MapGIS 10 (a geographic information system platform for managing and analyzing spatial data) and IGServer (a MapGIS software package for map service publication) are integrated into this layer [49]. In addition, the Hadoop ecosystem is integrated into this layer to provide storage and processing for unstructured data. In the GSISSP, a distributed Hadoop environment is deployed on each GSISSP node. All the primeval data are stored in an HDFS [51] (Hadoop Distributed File System), which can provide more reliable backup storage and distribute various access points over different servers. This system plays a key role in backing up original files. ZooKeeper [52] is utilized to coordinate and manage multiple Hadoop servers. To improve the access and retrieval efficiency, unstructured geological content is reorganized as "content items", where each content item contains multivariate characters, such as thematic features, spatial features, and temporal features. In GSISSP, HBase [53] provides storage support for content items.

The discovery and integration layer provides a node management method for gathering resources distributed in a cloud environment, and this layer describes the hardware resources, software resources and data resources universally. In the GSISSP, to achieve geological big-data integration, mining and knowledge-driven discovery, certain semantic technologies are integrated into the discovery and integration layer. The conceptual model between the terminologies is described by the geological domain ontology, and the relationships between the terminologies are expressed by an RDF (resource description framework), which defines the structure of the domain knowledge. The GSISSP integrates the geological thematic ontology, the geological temporal ontology and the toponymy ontology used to establish a basic knowledge-driven discovery environment. Combined with some semantic web rules predefined by domain experts, inference machines can be used to help identify new relationships that are not explicitly presented in the geological domain ontology. These new relationships are gathered to build a geological knowledge graph. Because this geological knowledge graph contains more abundant relationships and knowledge, information mining and discovery operations are launched. To discover spatial-related information, geospatial workflow services are integrated into this layer, and geospatial workflows can be executed automatically with a specific configuration file that is dynamically generated according to the SWRL and the retrieval framework ontology. In addition, certain data processing tools are integrated into this layer to increase the convenience of the content discovery. These tools are mainly used to reorganize the original data, extract content features, associate content items and batch upload data. Moreover, data-mining libraries (Mahout, PLDA, Nature Language Toolkit, etc.) can be gradually integrated in the future.

3. Organization of Multiple-Classification Content Based on Geologic Domain Ontology

3.1. Organization Patterns of Complex Geological Unstructured Content

Diversified and fragmented unstructured data from geological surveys are the most representational geological survey results. For example, a 1:250,000 regional geological survey report of Xigaze contains 308 different documents, and these documents are in different formats, such as .doc, .pdf, .jpg, and .tiff. The content of these documents includes geological maps, analysis reports and images. When managing these data for a certain region, all the files are assembled and defined as a geological archive that includes the concrete content and data features of a geological archive, as shown in Table 1:

Table 1. Concrete content and data features of a geological archive.

Document Name	Content Description	Count	Data Type	Data Features
Achievement reports	Regional geological survey theme reports (origin, evolution, working methods, etc.), results summaries	4	[a] doc	Unstructured
Achievement Illustrations	Geological maps, mineral maps, environmental geological maps	3	[b] gis	Structured
Acceptance Documentation	Final evaluation result reports, mid-term evaluation result reports, wild acceptance result reports	9	doc	Unstructured
Field Book	Wild route record, measured profile record	78	[c] xml, [d] jpg, doc, gis	Unstructured
Editorial Images	Field draft, comprehensive draft, factual material, primitive maps for compilation	31	jpg, gis	Unstructured

Table 1. *Cont.*

Document Name	Content Description	Count	Data Type	Data Features
Image and Interpretation	Remote sensing imagery, aerial survey interpretations	11	gis, [e] tiff, jpg	Unstructured
Geological Section	Geological section tables, stratum section column	54	gis, doc, [f] xls	Unstructured
Specimens	Ore mineral, spectrum, silicate, tombarthite	35	doc, gis	Structured
Quality Check	Geological data quality check card (geotraverse, section, etc.), raw material inspection records, raw material inspections	22	doc, xls	Structured
Measurement Report	Rock authentication reports, fossil authentication reports, tombarthite analysis report	21	doc, xls	Unstructured, Structured
Photos	Geological pictures	5	jpg	Unstructured
Designing Files	Overall design, geological mineral draft, project design, compilation note	7	doc, gis	Unstructured, Structured

[a] doc: Word document, [b] gis: MapGIS spatial data (includes the .wp, .wl, .wt, and .mpj formats), [c] xml: Extendable marked language file, [d] jpg: Pictures in .jpg format, [e] tiff: Remote sensing image, [f] xls: Excel document.

The information and knowledge contained in complex unstructured data are not expressed in the same way as traditional relational data. Instead, most of these data are included in the nature of unstructured text. Thus, building knowledge and feature libraries based on the original geological content is pivotal to effectively expressing information and represents the foundation to achieving knowledge discovery. The purpose of establishing a knowledge content library model is to reconstruct knowledge attributes or fragments based on the principle of not losing information. Thus, the original content data must be reorganized and described in as simple terms as possible to facilitate the virtual access of the data descriptions and the discovery of knowledge included in unstructured content. A knowledge content library model seeks to achieve descriptive and structural modeling based on clear semantics and efficient organization to promote knowledge integration and share and reuse information. In this paper, we aim to apply multiple classifications to unstructured content using various methods, such as content splitting, feature extraction, and information reorganization, to reconstruct unstructured geological data and store them in a NoSQL type database, which is more suitable for managing unstructured data. Because Hadoop ecosystem technologies have been merged into the GSISSP, HDFS and HBase are selected as the file system and database, respectively, to store geological unstructured content.

After storing the original geological documents in HDFS, our experimental process reorganizes and re-describes the original content to render the analysis closer to the data's original meaning and identify hidden knowledge without losing information. Based on the features of HBase, we can reorganize the original geological documents, store commonly used information in HBase as "content items" (forming a "basis content database"), organize unstructured data to render them suitable for data mining, and provide quick information services. The process of establishing the Basic_Content table is shown in Figure 2.

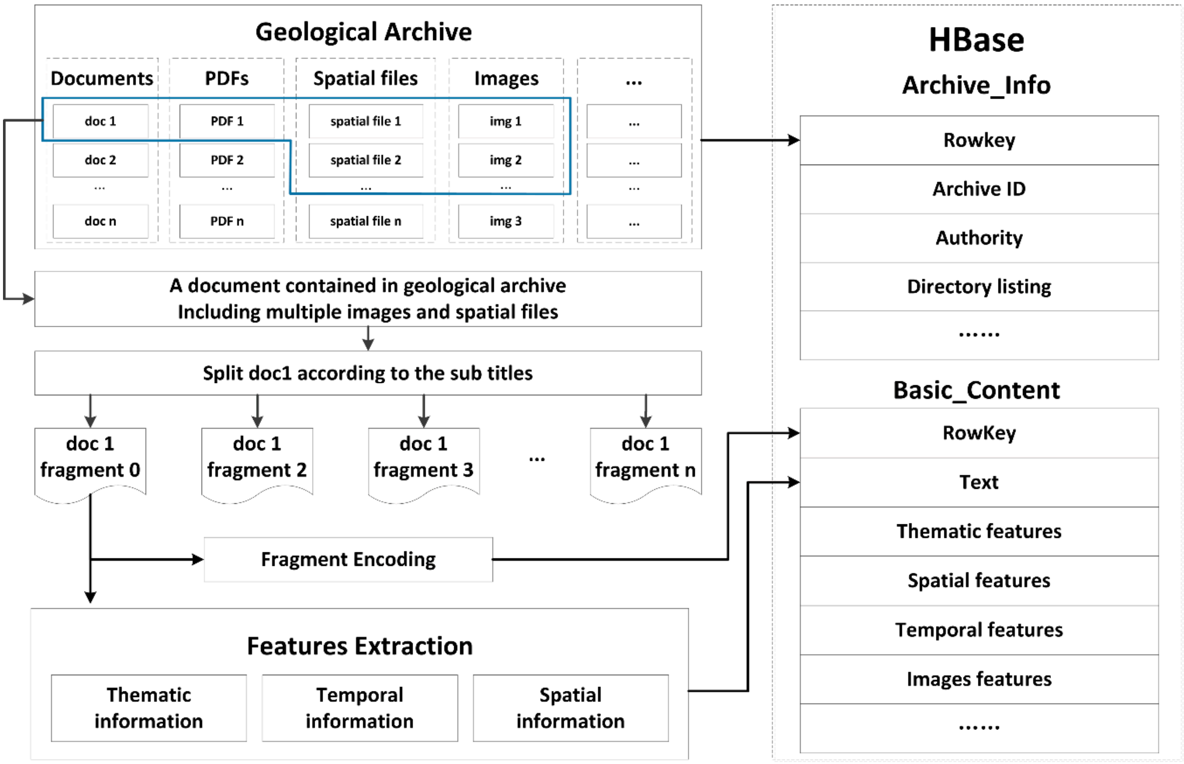


Figure 2. Organization of a geological archive.

Table 1 shows that a geological archive contains hundreds of documents alongside spatial data, pictures and spreadsheets, and each document usually references many illustrations and tables and corresponds to spatial data files. A directory listing in each geological archive records the metadata of all the included documents, pictures and spatial data. Generally, geological information services mainly focus on text, images, spatial data and temporal data. During the content-item-organization procedure, we extract the text, images, spatial information, time information and other common characteristics from geological archives and build an HBase table called "Basic_Content" to store these content items. As Figure 2 shows, each document in a geological archive is split into fragments, and the splitting process should be consistent with certain rules to maintain certain knowledge relationships, or the contents of each fragment should describe a certain subject. Thus, subtitles of documents are selected as the split point. The benefit of this strategy is that the main line of each fragment must revolve around a particular topic; therefore, our process is more convenient for extracting features. After splitting the document, three types of characteristics must be extracted from each fragment: geological thematic features, geological spatial features and geological temporal features. Thematic features describe the core content of the fragments, spatial features describe the spatial locations and attributes that are related to the fragment, and temporal features describe the geological time of the fragment. After extracting the thematic features, certain geological concepts regarding the fragment can be obtained. After extracting the spatial features, spatial location information (such as coordinates and map documents) or semantic spatial information (such as place names or metallogenic belts) can be obtained. After extracting the temporal features, the geological temporal terms of the fragment can be obtained. The extraction of these characteristics is mainly based on statistics, ontologies and semantic similarity theories. After the thematic, spatial and temporal characteristics are extracted, all the images in the fragment are extracted, which facilitates retrieval because the directory listing records the names and relative paths of all the image files. Thus, we can quickly determine which image is included in each fragment merely by matching strings.

After extracting the information from a fragment, all of the features related to this fragment can be organized as content items and stored in the HBase. As Figure 2 shows, the content items mainly include the following information: texts of fragments, thematic features, spatial information (e.g., coordinates and place names), corresponding map documents, temporal features, and all the image files. These pieces of information are stored in the Basic_Content table as a record. Each content item receives a unique code as the row key of the record; for example, in Table Basic_Content, the row key consists of a geological document's MD5 code and the fragment offset. In this way, the fragments from the same document can be stored together in order, and the fragments from different documents can be globally hashed to different servers, which will help improve the searching efficiency.

3.2. Storage Pattern of Complex Geological Unstructured Content

HBase retrieves a certain record via the key values, and well-designed row keys and column keys can significantly affect the efficiency of accessing data [54]. We designed two tables to store reorganized information: Table Archive_Info, which is used to store the metadata of geological original information, and Table Basic_Content, which is used to store content items. Details of Table Archive_Info's structure are provided in Table 2, and details of Table Basic_Content's structure are provided in Table 3. As Tables 2 and 3 show, "basic_info" and "feature_info" are two column families of Tables 2 and 3, respectively. Each column family contains more than one column to store different fields.

Table 2. Structure of Table Archive_Info.

	Row Key: MD5 of Geological Archive		
		Column_key	*Column name*
Table name: Archive_Info	Column family0: basic_info	arch_id	*Archive id*
		create_time	*Upload time*
		authority	*Access authority*
		user	*User name*
		size	*File size*
		name	*Archive name*
		title_ml_[a] [FilePath]	*[Document title]*

[a] [FilePath] represents the relative path of a certain file.

Table 3. Structure of Table Basic_Content.

	Row Key: MD5 of document + offset of fragment		
		Column_key	Column name
Table name: Basic_Content	Column family0: feature_info	frag_content	*Fragment content*
		theme_features	*Thematic feature concepts*
		time_features	*Temporal feature concepts*
		coordinate_info	*Coordinates*
		geo_name	*Toponymy*
		map_doc_[a] [FilePath]	*Source file of map document*
		breviary_img_[FilePath]	*[Thumbnail]*
		original_img_[FilePath]	*[Image original file]*

[a] [FilePath] represents the relative path of a certain file.

A fragment may contain more than one image or map document. In Basic_Content, not all of the images or map documents are stored in one column; instead, each image or map document will be stored as a single key value. For spatial data, HBase stores only the map document file (.mpj file format) and not the original spatial data (.wp, .wl, or .wt formats). Because the original spatial data are organized in HDFS according to the directory structure, the map document file records the relative positions of the original spatial data. Thus, when a map document is found, the corresponding spatial data can quickly be discovered in HDFS and published as a map service on IGServer.

4. Fragmented and Diversified Content Discovery

We establish a framework to help make the geological discovery system more knowledgeable. A retrieval frame ontology is designed based on ontology-related theories to achieve geospatial problem parsing and workflow building. We extract multiple features for each content item and link the content item with a specific geological ontology to improve the information accessibility of geological data. Thus, a bridge that joins the data and the framework is built.

4.1. Question-Oriented Content Retrieval Framework

The question-oriented content retrieval framework is an ontology-based framework to achieve geological knowledge and information discovery. In this framework, a specific class of geological question can be submitted according to the predefined SWRL, and the related spatial analysis operations can be semantically discovered. Then, a conceptual workflow is established automatically. The conceptual workflow can be executed with the spatial analysis services and work flow services in IGServer. The structure of the content discovery framework is illustrated in Figure 3. The entire framework is built on the resource layer, discovery and integration layer and application layer. The resource layer provides multiple spatial data, unstructured data, data services, spatial analysis services, and related services. The discovery and integration layer mainly combines ontologies, SWRL and semantic technologies to resolve the content discovery and work-flow-building problems. The application layer provides a retrieval portal and some visualization tools to receive geological problems and render analytical results.

Ontologies and source data must be organized into the system before using the framework:

a. Domain experts build and adjust the geological thematic ontology, geological temporal ontology and retrieval frame ontology;
b. We convert the original unstructured data to content items and extract the features of each item, and then we reorganize the content items to HBase and re-store the spatial data to HDFS.

For end users, the work process of this framework is as follows:

1. Submit a specific class of a geological question according to the retrieval portal, for example, the volcanic activity in Xinjiang, China;
2. Key information is extracted from the proposed question, such as the question type, target area, or target theme;
3. Find the corresponding rules from the predefined SWRL database according to the question type;
4. The SWRL rules are input into the inference machine;
5. The inference machine and SWRL rules bind with the geological domain ontology to start thematic reasoning and discovery;
6. If related content indices are discovered from the ontology, then the source unstructured data would be retrieved from HBase; if spatial data indices are discovered, then the source spatial data would be selected from HDFS and added to the GIS platform to facilitate the spatial analytical work. After related map documents or unstructured content are retrieved from the data source, the map document would be published as a map service on the IGServer platform;
7. Newly published map services are associated with the Web Service Ontology;

8. The inference machine and SWRL rules bind with the retrieval frame ontology to start the work flow reasoning and discovery, and then the conceptual work flow is built and expressed in a specific format;
9. The conceptual workflow is transmitted to the workflow engine and the spatial analytical work is initiated;
10. The spatial analytical results are transferred to the visualization tools in the application layer;
11. Visualization tools render the original result and show the final analytical result.

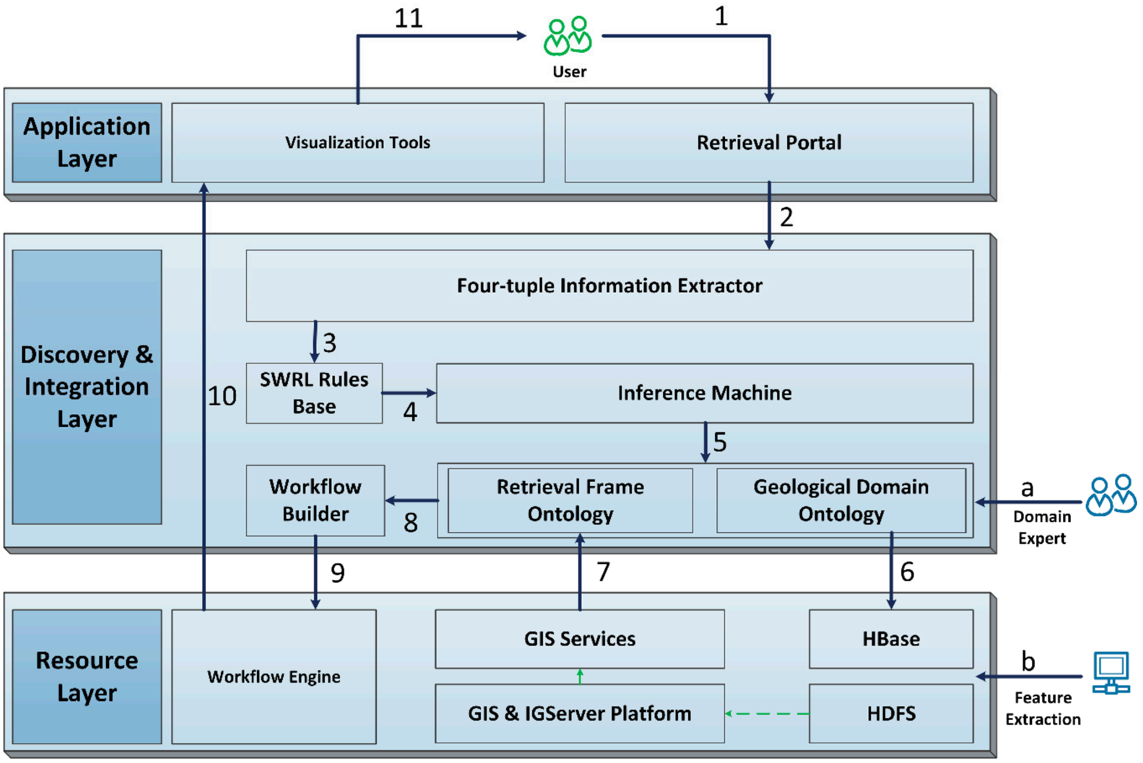


Figure 3. Question-oriented content retrieval framework.

4.2. Ontology Design

4.2.1. Geological Domain Ontology

Although geological data are expressed in a variety of forms, these data usually contain rich information. Thus, extracting information from massive high-dimensional unstructured data and associating spatiotemporal and thematic information are the keys to knowledge mining. To address these issues, we utilize ontological theories in the GSISSP to assist in resolving problems. The concept of ontology originated from the field of philosophy, which is used to explain the nature of existence [55]. From the perspective of information science, ontology is a conceptual model that describes the term and the relationship between terms. This concept is a clear specification of the conceptual model [56,57]. On the one hand, ontology limits the term set, so we must use a common recognition of a set of words. On the other hand, ontology defines the upper and lower relationships between the terms. Ontologies can be used to identify and correlate the knowledge that corresponds to the information concept to realize the explicit semantics of information content [32,58]. The geological ontologies used in this paper consist of a geological thematic ontology, toponymy ontology and geological temporal ontology (Figure 4).

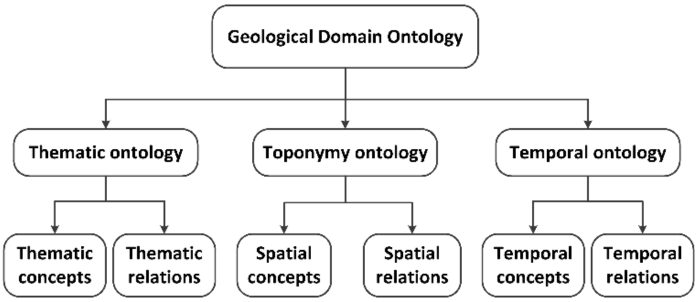


Figure 4. Structure of a geological domain ontology.

A geological ontology is a domain ontology that describes the knowledge of a geological field. This ontology provides the formal concepts, interrelationships, characteristics, and laws of the geological field. For geological content retrieval and discovery systems, geological domain ontologies could help eliminate the ambiguity surrounding geological concepts or terms and provide a common understanding of geological knowledge and a clear definition of relationships between geological terms at different levels [59,60]. In our previous research, geological experts constructed a geological thematic ontology that includes 22 categories (e.g., rock, stratum, and geological structure) and more than 49,520 geological professional terms and concepts; this ontology provides a detailed hyponymy classification of the geology [61,62]. A portion of this geological thematic ontology is shown in Figure 5. Moreover, a geological temporal ontology has been constructed by geological experts, and it includes 737 geological temporal-related terms (e.g., geological age, geological movement, glacial period, rain period, and biological evolution stage). This ontology also includes the hyponymy between terms, and a portion of the geological temporal ontology is shown in Figure 6. In this study, we consider the spatial attributes in the geological domain using the place names database and the relationships between the place names. Thus, we built a toponymy ontology that is accurate to the village and street level (as shown in Figure 7) and that combines spatial data to provide location-related information.

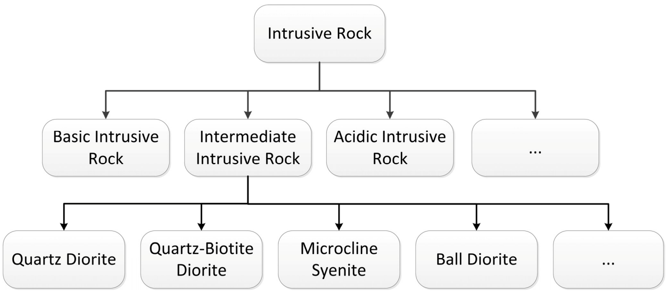


Figure 5. Sample of a geological thematic ontology.

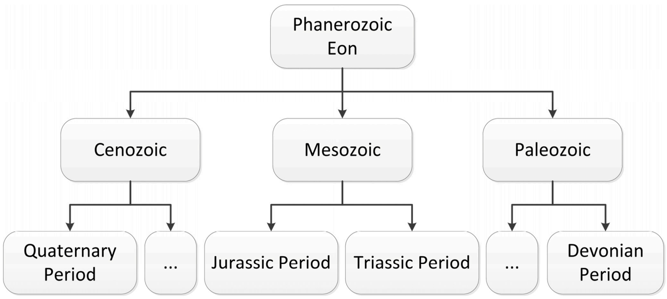


Figure 6. Sample of a geological temporal ontology.

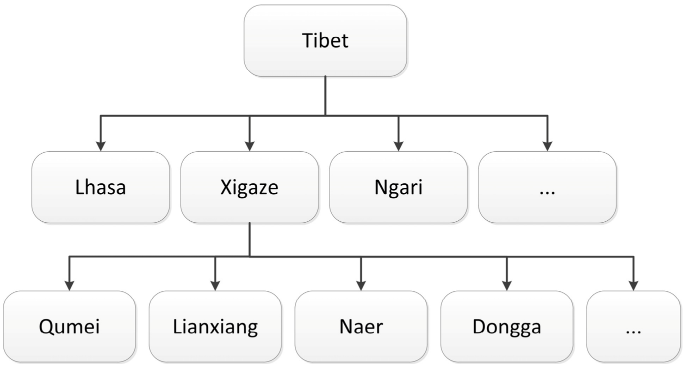


Figure 7. Sample of a toponymy ontology.

4.2.2. Retrieval Frame Ontology

Geological problems usually involve spatial analysis operations. To provide accurate spatial information, the GSISSP must consider how to query spatial data and text content related to target problems and determine how to intelligently complete the related spatial analysis process and return the results. Furthermore, the entire process of spatial analysis should be transparent to the users. To achieve an intelligent content discovery service, spatial analysis service, and the visualization of the results based on previous work, this paper designs a retrieval frame ontology by binding the SWRL rules and web services. The retrieval framework can convert a specific class of geological problem into a workflow process, and spatial analyses, content discovery and other operations can be executed using a workflow engine. The structure of our retrieval frame ontology is shown in Figure 8.

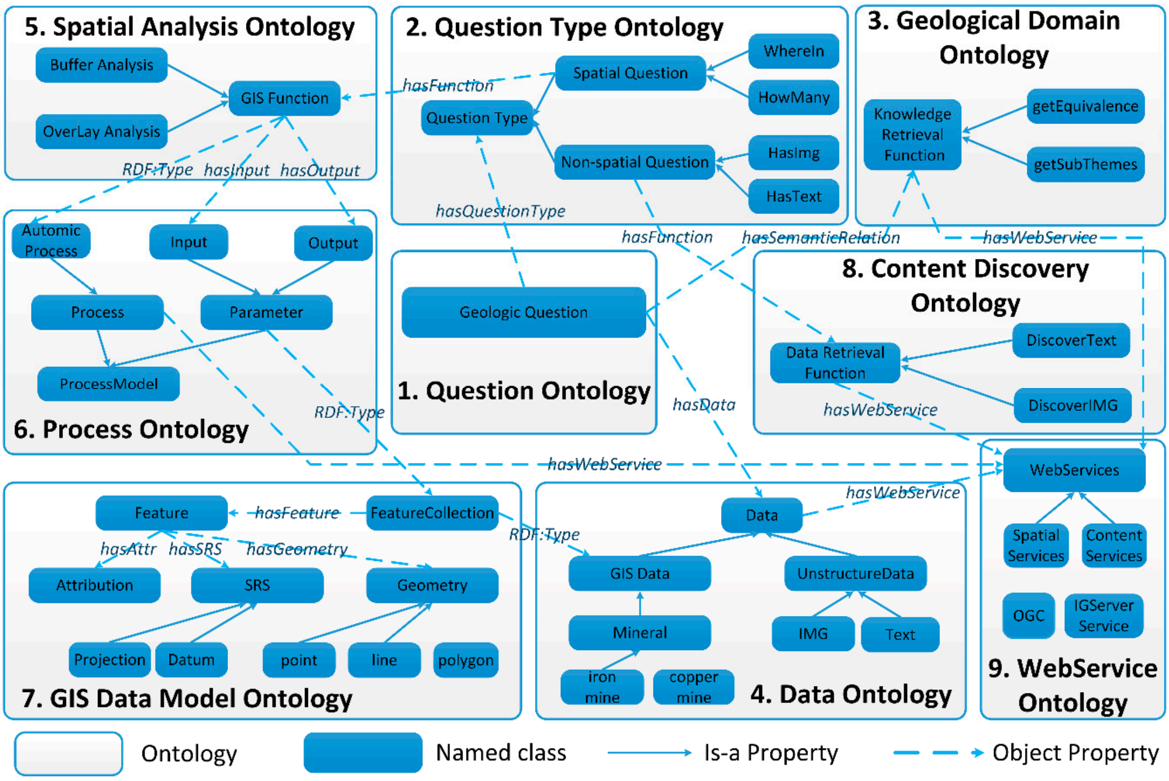


Figure 8. Structure of the content retrieval frame ontology.

Figure 8 shows a specialized ontology that is designed to solve certain geospatial problems. The entire framework consists of sub-module ontologies that are each incorporated with SWRL rules; thus, the framework has an inferential capability. By binding predefined SWRL rules to the inference engine, automatic content discovery can be achieved. Each inference rule implies the necessary conditions and branching conditions for every step of the operation. Hence, according to the inference rules, the whole concept workflow process can be established. After rewriting the concept workflow in a specific workflow engine format, the concept workflow can be executed by the workflow engine. In this study, MapGIS 10 is used to provide the workflow engine and IGServer is used to provide workflow-related web services. Throughout the entire framework, different types of ontologies are connected through object properties, and the function of each sub-module ontology is as follows:

1. Question Ontology:
   The question ontology defines the conceptual model of the geological question. All the geological questions must contain two elements: the question type and the target content. The question ontology is associated with the question type ontology and data ontology; the objective is to verify both the geological question type and the related data content that are queried by the users. To obtain knowledgeable or semantic information that is related to the target content, the question ontology is associated with the geological domain ontology to more comprehensively render the information discovery.
2. Question Type Ontology:
   The question type ontology describes the concrete question types. Two main types of questions exist in the current framework: spatial-related questions and non-spatial-related questions. Different spatial analysis operations must be employed to answer a spatial-related question according to the different questions. For example, *whatNear* and whereIn are two spatial-related questions, and the *whatNear* type of question can be semantically expressed as "what is near a target area". This type of question semantically implies a buffer analysis operation. The whereIn type of question can be semantically expressed as "where are the objects in the target area". This type of question semantically implies an overlay analysis operation. Spatial questions are associated with the spatial analysis ontology to discover the related spatial analysis operations. Non-spatial-related questions are different from spatial-related questions. To answer these questions, the target content must be retrieved from the database without undertaking a spatial analysis. Consequently, non-spatial questions do not have to be associated with the spatial analysis ontology.
3. Geological Domain Ontology:
   The geological domain ontology provides the domain knowledge and information to achieve knowledgeable and semantic discoveries in the retrieval framework. According to the geological domain ontology, the keywords that are contained in the retrieval question can be obtained. In addition, the equivalent words and hyponyms of these keywords can also be obtained according to the relations that are contained in the geological domain ontology. During the retrieval, equivalent words and hyponym words will also be retrieved, which could guarantee comprehensiveness with respect to the semantics and knowledge. In the geological domain ontology, knowledgeable information is obtained by invoking a related web service, so the geological domain ontology is associated with the web service ontology.
4. Data Ontology:
   The data ontology describes the involved data content contained in the geological questions. There are two types of data in the framework: spatial data and unstructured data. Although spatial data contain multiple types of geological thematic data, unstructured data contain various geological pictures and text. All the data access operations are implemented by a web service, so the data ontology is associated with the web service ontology.

5. Spatial Analysis Ontology:
   The spatial analysis ontology describes all of the common spatial analysis functions, and it semantically annotates the associated input, output and algorithm model of these spatial analysis functions. The spatial analysis ontology is connected to the process ontology.
6. Process Ontology:
   The process ontology describes the execution standard of the spatial analysis web services and annotates the input and output of these services. This ontology is connected to the related spatial analysis services.
7. GIS Data Model Ontology:
   The GIS data model ontology describes the organization model of the spatial data. All of the spatial processing work will finally be concentrated on the feature processing, as each spatial feature contains attributes, references and geometries. Here, we utilize the spatial data organization model of MapGIS 10, and the GIS data model ontology inherits the ontology in the GSICCP [49].
8. Content Discovery Ontology:
   The content discovery ontology describes the process function that does not involve spatial data, and it currently includes image discovery and text discovery functions. This ontology is implemented by a web service and is connected to the related web services.
9. Web Service Ontology:
   The web service ontology describes all of the web services in the GSICCP, including spatial-related web services and non-spatial-related web services. Although the spatial-related web services mainly consist of the OGC service (WPS, WFS, etc.) and the IGServer service (the map document service, tile map service, vector layer service, etc.), non-spatial-related web services mainly consist of unstructured content discovery services and some data-mining services.

The classes of each ontology in Figure 8 are connected to other classes according to a series of relationships (*has** format). This type of relationship is defined as an object property in the ontology, which reflects the association between the objects in the ontology. Another type of relationship that corresponds to the object property is the data property, which reflects the attributes of the object itself. In this retrieval frame ontology, the parsing of problems and building of the workflow are achieved based on the object properties. For example, the GIS Function class in the Spatial Analysis Ontology (5th in Figure 8) contains some object properties: the *hasInput* property is associated with the Input class of the Process Ontology, and the *hasOutput* property is associated with the Output class of the Process Ontology and is an individual of (rdf: type) the Automatic Process. The Input class and Output class are the subclasses of the Parameter class, which is an individual of the FeatureCollection class. All the individuals of FeatureCollection are spatial data that are expressed in the GIS data model. According to the retrieval frame ontology, a correct completion of a spatial analysis operation requires some spatial data as inputs, and the results are returned to the spatial data. Any individual in the GIS Function is an automatic operation and thus clearly expresses the function of each class and its individuals.

For instance, consider the problem "how many freshwater lakes in China". The process of automatically building a workflow according to the framework is illustrated in Figure 9. When a question is submitted, a semantic parser explains the question in the form of a 4-tuple, which is designed as <question_type> <target_theme> <spatial_relationship> <target_area>. Hence, the target question is transformed into <HowMany> <freshwater lakes> <In> <China> and sent into the system. The <question_type> and <spatial_relationship> can determine the type of problem and the spatial operations. The <target_theme> and <target_area> can determine the required data for answering the question. Different classes of geospatial problems have their own spatial operation methods and processes, so different types of problems must choose different spatial analysis functions. SWRL was used to determine the spatial analysis functions and sequence under different problems. In SWRL, the input constraints and outputs that meet all the constraints are defined by analyzing the SWRL rules for different types of questions. We could determine the prerequisites and final results to resolve the problem, thus laying the foundation for the automatic building of the workflow. In the example question, the related SWRLs of a "HowMany" question are defined as follows:

- Rule 1: define a formal question composition, which contains a question type and required data.
- Rule 2: define a HowMany question; a question that is marked with "HowMany" could be regarded as a "HowMany" question.
- Rule 3: define the execution sequence of the HowMany question, which requires an overlay and spatial query operation; the spatial query operation follows the overlay operation.
- Rule 4: define the function's connections; the output of the previous function is the input of its subsequent functions.
- Rule 5: define the requirement of identical data; two identical spatial data set must have identical geometries and spatial reference systems.

```
Rule1:
hasQuestionType(?type) ∧ hasData(?type) →
Question(?type)
Rule2:
HasQuestionType("How Many") ∧ hasData?(?type) →
HowManyType(?type)
Rule3:
HowManyType(?type) ∧ Overlay(?overlay) ∧ SpatialQuery(?spatialQuery) ∧
hasGISFunction(?type, ?overlay) ∧ hasGISFunction(?type, ?spatialQuery) ∧
hasNextFunction(?overlay, ?spatialQuery) →
howMany(?type)
Rule4:
GISFunction(?func1) ∧ Output(?out1) ∧ hasOutput(?func1, ?out1) ∧
GISFunction(?func2) ∧ Input(?in1) ∧ hasInput(?func2, ?in1) ∧
hasSameData(?out1, ?in1) →
hasNextFunction(?func1, ?func2)
Rule5:
FeatureCollection(?data1) ∧ FeatureCollection(?data2) ∧ Geometry(?geo1) ∧
Geometry(?geo2) ∧ hasGeometryType(?data1, ?geo1) ∧ hasGeometryType(?data2, ?geo2) ∧
hasSRS(?geo1, ?srs1) ∧ hasSRS(?geo2, ?srs2) ∧ sameAs(?geo1, ?geo2)
∧ sameAs(?srs1, ?srs2) →
hasSameData(?data1, ?data2)
```

SWRL shows that spatial overlay analysis and a spatial query function are required to solve "HowMany" questions. The GIS Function class is questioned according to the object property *hasFunction* of the Spatial Question to determine whether spatial overlay analysis and a spatial query function exist. If functions are found, the required inputs and target output can be determined for each concrete spatial analysis function according to the *hasInput* and *hasOutput* properties to filter opposite functions. After finding all the appropriate spatial analysis tools, we could invoke the specific spatial analysis function through the web service. We could determine the context of the spatial analysis according to the sequence that was defined in the SWRL rule. We could determine the appropriate data or parameters and identify the entire workflow according to the input and output conditions of each space operation. Finally, we could describe the entire workflow process with the specific workflow language and start the engine to execute the workflow.

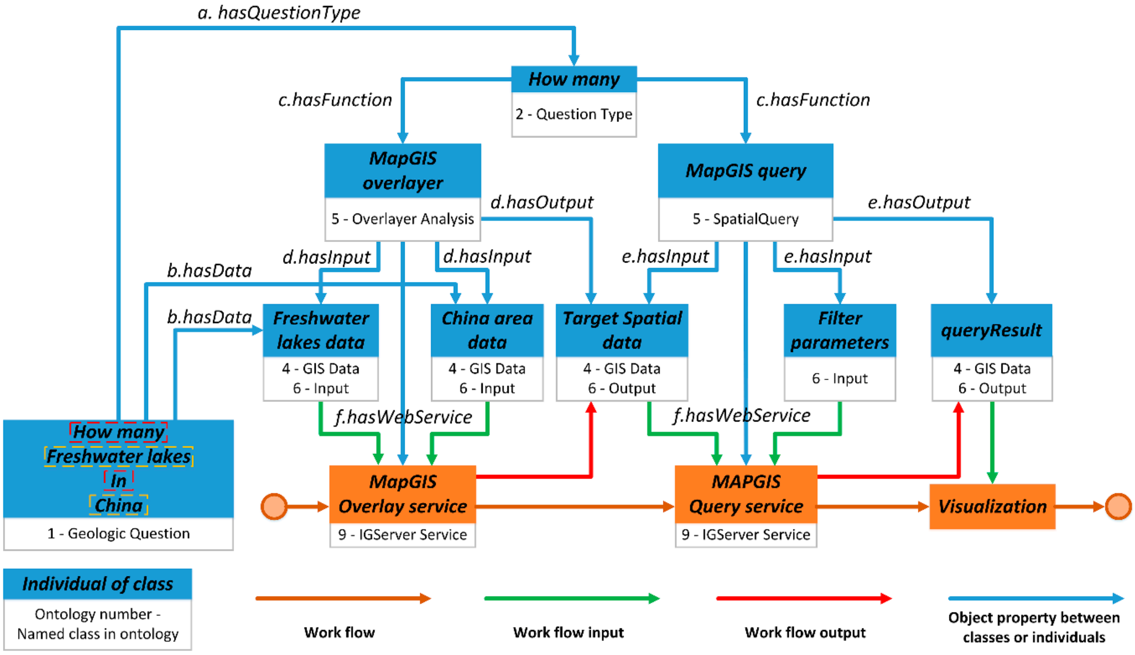


Figure 9. Process of automatically building a workflow.

4.3. Feature Extraction Based on the Geological Domain Ontology

4.3.1. Thematic Feature Extraction

This paper selects the sentence in each fragment as the semantic annotation unit. Based on the geological thematic ontology, the model in this paper extracts implicit information from the text. A geological document may contain multiple topics, so we may experience a partial fragment loss problem if we utilize only statistical methods to extract pieces of knowledge in accordance with the paragraph weight. Therefore, we used sentences as the basic unit for information extraction. Under the guidance of the document theme and concept similarity theories, we use statistical methods and heuristic rules to extract the key sentences from the text as the abstracts of the fragments. Finally, we extract all of the geological terms from the abstract. The specific steps are provided below.

First, the text is split into multiple individual words, and certain words that are not relevant to the subject knowledge (for example, prepositions, stop words, and function words) are eliminated. Only the key terms that are described in the geological ontology are preserved. After eliminating the irrelevant terms from the sentence, the sentence is expressed in the form of a term feature vector. Certain definitions are listed below.

Definition 1. *Assume that the fragment consists of sentences $S_1, S_2, \ldots, S_m$, where sentence S_i consists of the term (or keyword) set $N_i = \{ N_{i1}, N_{i2}, \ldots, N_{ik} \}$. Then, the term set in the fragment can be expressed as follows:*

$$N = \bigcup_{t=1}^{m} N_t = \{ N_1, N_2, \ldots, N_m \} = \{ T_1, T_2, \ldots, T_n \} \quad (1)$$

T_n is the term identified from the fragment text that can be used to represent a thematic concept.

Definition 2. *The weight of term T_i in a fragment can be expressed as follows:*

$$W_i = F_i \times log_2\left(\frac{n_i}{m} + 1 \right) \quad (0 < n_i \leq m) \quad (2)$$

where F_i is the frequency value of term T_i in the fragment; m is the count of all the sentences in the fragment; n_i is the count of the sentences that contain the term T_i; and n_i/m reflects the coverage rate of term T_i in the fragment.

Definition 3. *Sentence S_i can be expressed in the form of a vector as $(STW_{i1}, STW_{i2}, \ldots, STW_{ij}, \ldots, STW_{in})$, where STW_{ij} is the weight of term T_j in sentence S_i and*

$$STW_{ij} = F_{ij} \times W_j \times \frac{n_j}{m} \quad (0 < n_j \leq m) \tag{3}$$

In the above equation, F_{ij} is the frequency value of term T_j in sentence S_i; W_j is the weight of term T_j in the fragment; n_j is the number of sentences that contain the term T_j; and m is the number of sentences in the fragment.

Definition 4. *The weight of sentence S_i is as follows:*

$$SW_i = STW_{i1} + STW_{i2} + \ldots + STW_{in} \tag{4}$$

Finally, all the sentences whose weights are over the threshold value are reserved and formed as an abstract of the fragment, and all the geological thematic concepts that emerged in the abstract are extracted as the thematic features of this fragment. Figure 10 illustrates the process of the thematic extraction approach; the left half of figure describes the input and output during the whole process and right half describes the pivotal processor of the whole flow.

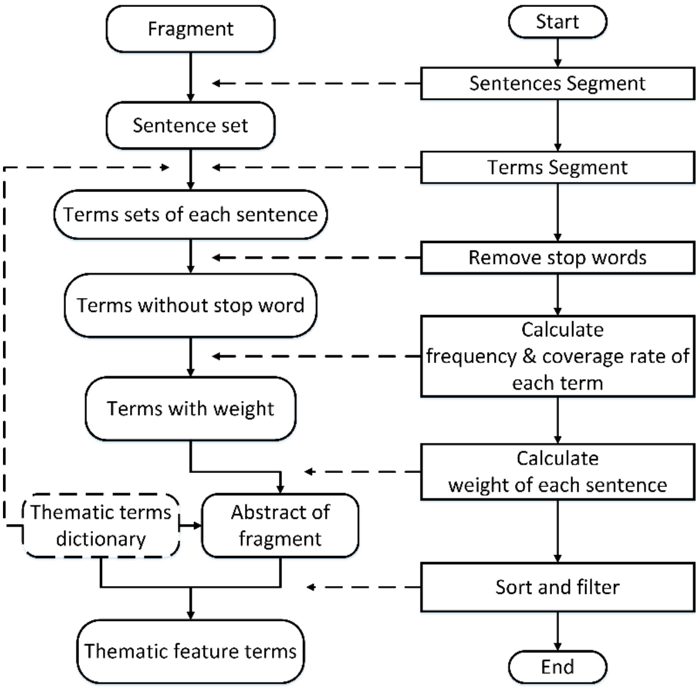


Figure 10. Process of geological thematic feature extraction.

4.3.2. Spatiotemporal Feature Extraction

The spatial information that is included in the geological survey results mainly contains some toponymy, coordinates, and spatial data. The toponymy and coordinates are mainly extracted from the text. Although the process of toponymy extraction is similar to the thematic feature extraction process, the toponymy ontology provides the guidance instead of the geological thematic ontology. Coordinate extraction would be easier because the coordinates are usually expressed in a specific form. By using regular expressions, the content of the coordinates can be quickly identified. Then, according to the specific spatial-information services provided by the GSISSP, the relevant toponymy and geological body information can be obtained rapidly.

The spatial files contained in geological archives generally have a specific format or suffix. Thus, when we serialize the geological contents with a specific data structure, we can determine whether the contents contain spatial data. The process of identifying spatial data is depicted in Figure 11. First, all the map documents are converted into the format of MapGIS 10; then, the geographic extent and annotation context are extracted and stored in the Basic_Content table, and each map document is stored in its own column in the Basic_Content table. IGServer provides visualizations for spatial data, and "spatial data processing" can automatically identify, convert and publish GIS data. HDFS stores the original spatial data file and acts as a data source for map services. The temporal features extraction process is similar to the thematic feature extraction process, although the geological temporal ontology will provide the guidance instead of the geological theme ontology.

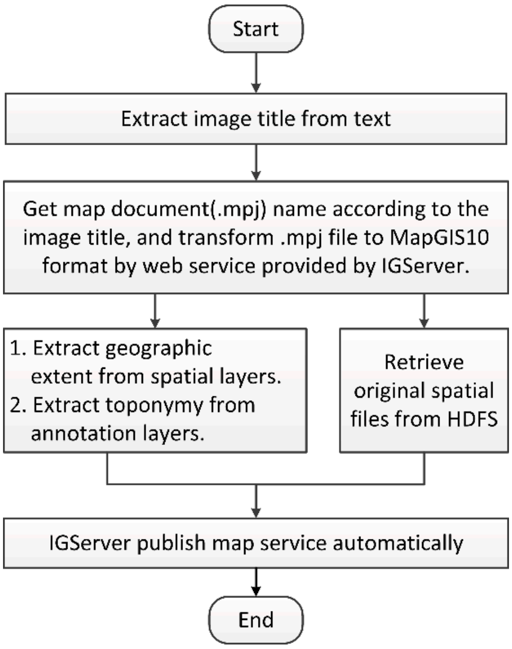

Start
Extract image title from text
Get map document(.mpj) name according to the image title, and transform .mpj file to MapGIS10 format by web service provided by IGServer.
1. Extract geographic extent from spatial layers.
2. Extract toponymy from annotation layers.
Retrieve original spatial files from HDFS
IGServer publish map service automatically
End

Figure 11. Process of spatial content discovery.

4.4. Multi-Feature Content Association

After organizing the unstructured fragments as content items, the content items in the HBase table can be formally expressed as a 4-tuple:

$$C = \{\ k, v, R_c, E\ \} \tag{5}$$

where C represents a content item, k is the key of content item C and has a unique value, and v is the text value of C such that one k maps to one v. Moreover, R_c represents the relationships between one content item C and another content item C', and these relationships mainly include the parent-child relationship and inclusion relationships. Using an object-oriented model, such a relationship can be expressed as a generalization, association, aggregation, composition or dependency. In addition, E represents multivariate features of C, which can be expressed as follows:

$$E = \{\ e_i | \forall e_i \in D_i, 0 \leq i \leq n\ \} \tag{6}$$

where e_i is a feature value of one dimension and D_i are the value ranges of the feature e_i. Theoretically, E is an n-dimensional feature vector, although each e_i belongs to only one D_i (e.g., the feature could be a thematic feature, temporal feature, or spatial feature). $D.$ is defined as the feature range:

$$D = \bigcup_{i=1}^{N} D_i \tag{7}$$

D meets the conditions $D_i \subset D, D_j \subset D, D_i \cap D_j = \varnothing$, which implies that D is the union of a plurality of disjoint feature fields. Therefore, the count of the sub-fields equals the count of the feature dimensions in a content item. A feature e_i that appears in multiple content items is defined as a flyweight, which is one of the ways in which we achieve the aggregation and classification of the content items. The composition and structure of the relationship between the content items is depicted in Figure 12. The upper half of Figure 12 is a schematic representation of the ontology: the ontology contains different named classes, and some associations exist between the named classes, for example, an upper or lower relationship (relationship between node A and node B), sibling relationship (relationship between node C and node D) and other complex relationships. These basic relationships constitute an ontology tree model. In an ontology, the "class" concept is at an abstract level: individuals of a class are the real content that is valuable to users. Currently, the individuals are organized as content items in HBase; therefore, we must associate the content items with specific named classes to achieve content discovery with assistance from the associations in the ontology. The lower half of Figure 12 shows the links of the named classes and individuals: the content items build the relationship with the named classes of thematic, spatial and temporal ontologies through their thematic, spatial and temporal features; the dotted lines with arrows in Figure 12 reflect these relationships. Corresponding to Equation (5), Rowkey uniquely identifies a content item (k in Equation (5)), the fragment text field contains all the unstructured data (v in Equation (5)), the relationships between the named classes indirectly reflect the associations between content items (R_c in Equation (5)), and all the feature fields in the content item represent n-dimensional features (E in Equation (5)). These relationships are stored in RDF format, which can be queried by SPARQL. All the associated content is distinguished to further identify content that is related to a thematic, location or temporal concept.

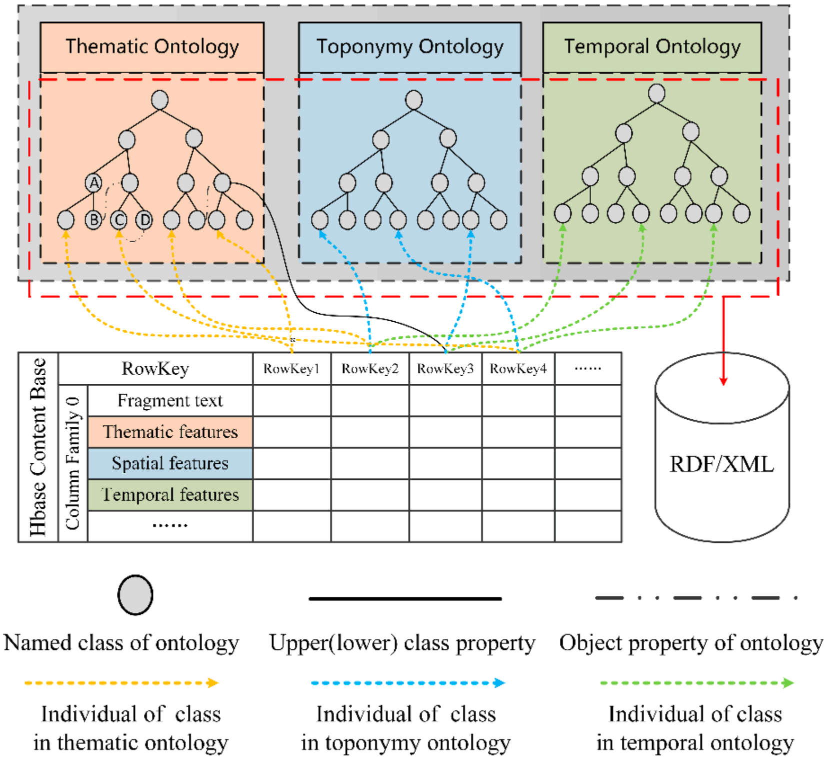


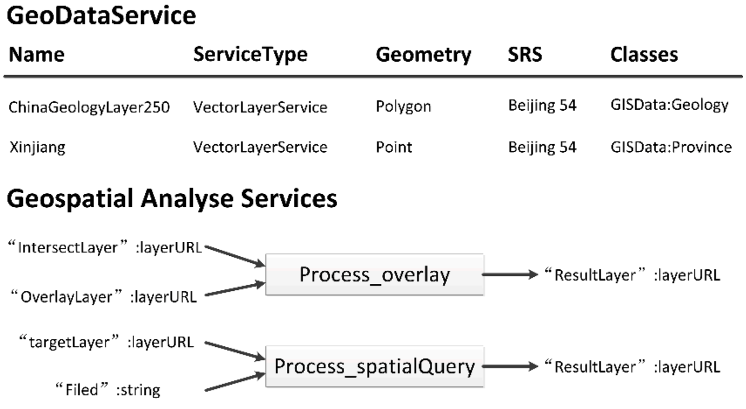
Figure 12. Associations of the content items.

5. Use Case

In our previous work, geological archives from Xinjiang were reorganized and stored in a Hadoop environment. Xinjiang is in the northwest region of China, which has intensive volcanic activity and several types of volcanoes, and it presents the most widely distributed volcanic activity in China. We selected "volcanic distribution in Xinjiang area of China" as a target question to illustrate the working mechanism of the entire framework. Thus, "volcanic distribution in Xinjiang area of China" can be classified as a *whereIn* type of question, which can be semantically expressed as "where is the volcanic activity in Xinjiang, China". For *whereIn* questions, the SWRL rules are defined as follows:

Rule1:
hasQuestionType("WhereIn") ∧ hasData(?type) →
WhereIn(?type)
Rule2:
WhereIn(?type) ∧ Overlay(?overlay) ∧ SpatialQuery(?spatialQuery) ∧
hasGISFunction(?type, ?overlay) ∧ hasGISFunction(?type, ?spatialQuery) ∧
hasNextFunction(?overlay, ?spatialQuery) →
whereIn(?type)

In the SWRL rules, the basic workflow of a *whereIn* question is predefined. These types of questions involve two categories of spatial operations: overlay operations and spatial query operations. First, the thematic data in the target area are obtained according to the overlay operation. Then, the spatial query operation is performed to obtain the thematic contents that meet the proposed criteria. Before we start the experiment, spatial data, spatial web services and unstructured geological documents were uploaded into the system in addition to semantic annotations that were previously added to the spatial data and spatial services. We associate the data and services with the geological domain ontology and retrieval frame ontology according to the semantic annotations (Figure 13). For unstructured geological documents, the fragments are reorganized and stored in HBase and associated with the geological domain ontology after the document splitting and feature extraction (previously described).



GeoDataService

Name	ServiceType	Geometry	SRS	Classes
ChinaGeologyLayer250	VectorLayerService	Polygon	Beijing 54	GISData:Geology
Xinjiang	VectorLayerService	Point	Beijing 54	GISData:Province

Geospatial Analyse Services

Figure 13. Spatial data services and spatial analytical services.

The target question "volcanic distribution in Xinjiang area of China" is submitted, transformed into <where><volcanic><In><Xinjiang>, and sent into the system. Then, the SWRL rules that correspond to the *whereIn* question type are parsed. We found that the overlay and spatial query analyses must be undertaken one by one because the output of the overlay is the input of the spatial query. By constructing SPARQL retrieval expressions and executing query operations, we discovered that the spatial services process_overlay and process_spatialQuery are associated in the ontology. In addition, according to the <target_theme> and <target_area>, the spatial data and unstructured content of the target area are retrieved from HBase and HDFS. When all of the data that meet the criteria have been found, the workflow services automatically begin. After entering the related data and invoking the target web services, the analytical results and all the unstructured content that is related to "volcanic in Xinjiang" are discovered from HBase (Figure 14). Finally, the spatial results in the maps and unstructured results are comprehensively displayed (Figure 15).

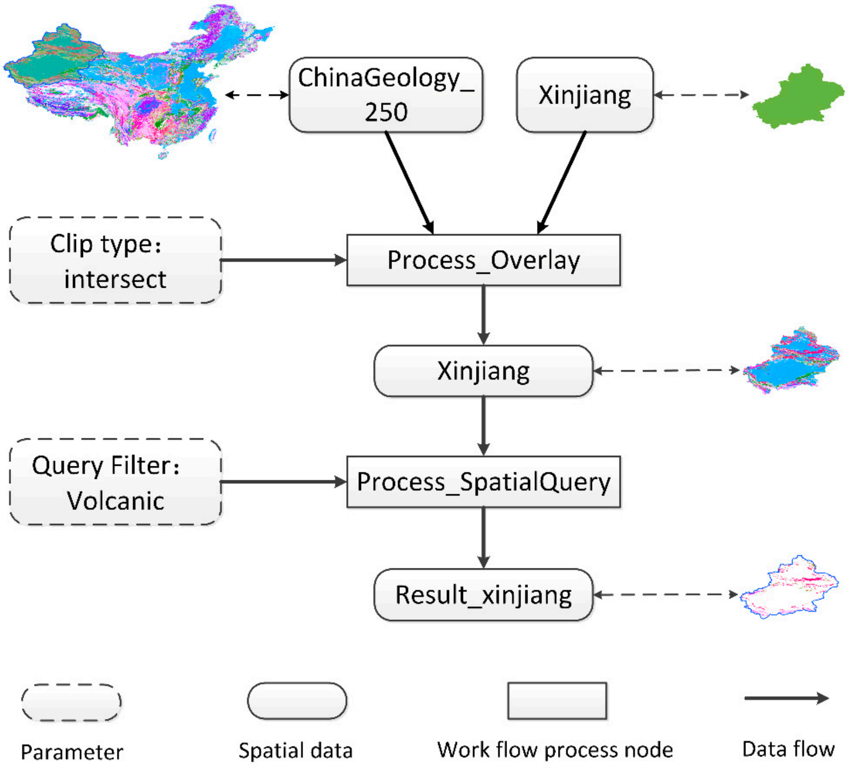
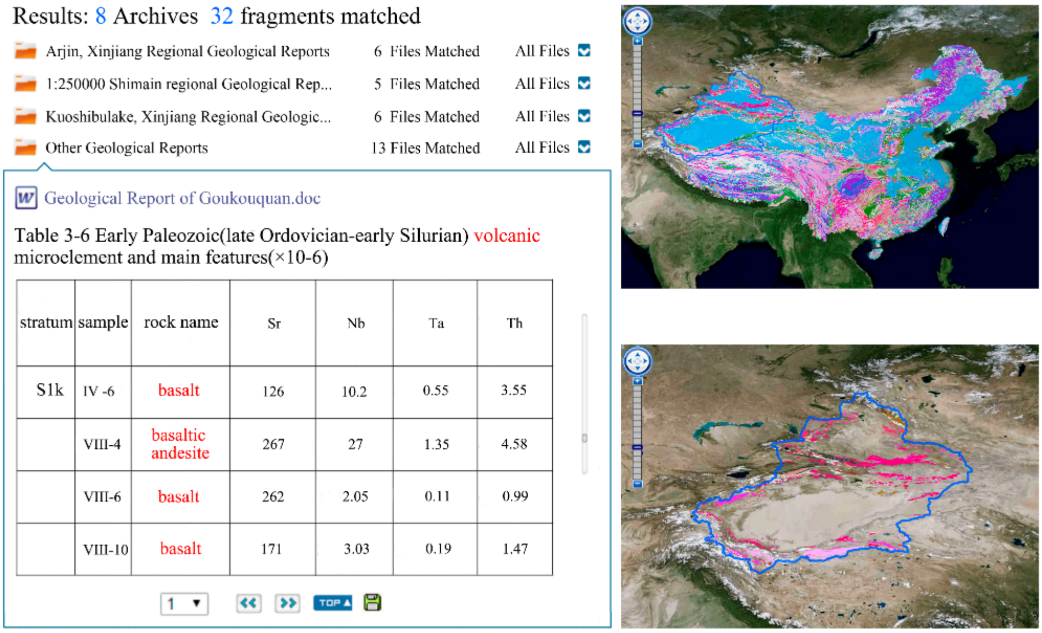


Figure 14. Process of workflow building.



Results: 8 Archives 32 fragments matched

Arjin, Xinjiang Regional Geological Reports	6 Files Matched	All Files
1:250000 Shimain regional Geological Rep...	5 Files Matched	All Files
Kuoshibulake, Xinjiang Regional Geologic...	6 Files Matched	All Files
Other Geological Reports	13 Files Matched	All Files

Geological Report of Goukouquan.doc

Table 3-6 Early Paleozoic(late Ordovician-early Silurian) volcanic microelement and main features(×10-6)

stratum	sample	rock name	Sr	Nb	Ta	Th
S1k	IV -6	basalt	126	10.2	0.55	3.55
	VIII-4	basaltic andesite	267	27	1.35	4.58
	VIII-6	basalt	262	2.05	0.11	0.99
	VIII-10	basalt	171	3.03	0.19	1.47

Figure 15. Display of the content retrieval result.

Figure 15 shows an integrated result from the query. The left half of the figure represents the unstructured data that were retrieved from HBase, and the right half of the figure shows the corresponding spatial analysis result, which is displayed in a map. Usually, the unstructured retrieval result contains text, tables, images, etc., which is different from keyword matching-based retrieval. We achieved the content retrieval with a geological ontology to promote the keyword matching-based retrieval to semantic retrieval. As Figure 15 shows, the system searched for the word "volcanic" from the database, but other concepts such as basalt and basaltic andesite were also returned because the term "volcanic" has a close relationship with "basalt" in geology. To measure the effect of semantic-based retrieval, we conducted a comparative test to verify the change in the precision rate and recall rate during content-item retrieval. Group 1 adopted the keyword matching-based method, which was implemented with Lucene (an open source full-text search framework that was developed by Apache), and Group 2 adopted our semantic-based method. We took a sample that contained 100 content items and artificially determined the relevance of the search term and each content item. The test results are shown in Table 4:

Table 4. Comparison of the retrieval results.

Group	Item	Method	[a] R	[b] H	[c] N	[d] S	[e] Precision	[f] Recall
1	Volcanic rocks	Lucene	51	39	48	100	76.47%	81.25%
2		Semantic	67	46	48	100	68.66%	95.83%
1	Metamorphic rocks	Lucene	42	26	33	100	61.90%	78.77%
2		Semantic	40	29	33	100	72.50%	87.88%
1	Arjin	Lucene	59	43	57	100	72.88%	75.44%
2		Semantic	74	56	57	100	75.68%	98.25%

[a] R returned content item; [b] H related content items of returned content items; [c] N related content item of sample; [d] S total amount of content items in sample; [e] Precision = H/R; [f] Recall = H/N.

6. Discussion

This paper is based on previous research results, but it introduces technologies related to big data and semantics into the GSICCP. Moreover, this paper promotes the following three aspects of the GSICCP: massive complex geological unstructured data organization, geological knowledge-based relationship construction and knowledge-driven geological content discovery. This paper constructed the GSISSP to enhance the information accessibility of the geological survey domain.

1. Massive unstructured complex geological data organization

   In this study, we aimed to characterize geological survey data (e.g., massive, complex, diversified, or unstructured) and enhanced the data-management methods with Hadoop ecosystem technologies. HDFS was utilized to store the original geological content. Instead of using a traditional file system or relational database, the original geological content in HDFS had multiple copies. Compared to traditional file-based storage, the use of the distributed architecture of the Hadoop system technology may have improved the security of the original data and improved the concurrent data access efficiency. In addition, massive geological unstructured data were split into fragments, and a NoSQL database was utilized to reorganize these fragmented contents. Split unstructured data could help reduce the complexity and heterogeneity of original geological data. Therefore, we fully exploited the NoSQL database features to efficiently manage the fragmented geological content, which improved the retrieval and computational efficiency of the geological unstructured data.

2. Constructing geological knowledge relationships

   In this paper, we addressed the thematic, spatial and temporal features of the geological domain, introduced a geological thematic ontology and geological temporal ontology, built a toponymy ontology, and extracted multi-dimensional features from geological unstructured data.

The geological domain problem usually contains spatiotemporal characteristics, and the extracted thematic, spatial and temporal features would be advantageous to geological content discovery work that considers spatiotemporal attributes. Moreover, associations between the fragments, spatial data and images were built based on the relationships that were represented in the ontologies and the extracted features, and these associations subsequently produced intelligent and semantically driven connections among the geological content. This set of associations built the foundation for knowledge-driven geological content discovery.

3. Knowledge-driven geological content discovery

   Many unstructured content retrieval studies have been performed in the past. However, in our previous work, we discovered geological content according to keywords in the text by extracting words from the text and building an index between the words and the content. When a query request is sent, the index of keywords is used to identify all the fragments that contain the keywords. However, this strategy has a huge gap: keyword-based searching ignores semantic and knowledge relationships. Thus, certain concepts that have the same semantics but are expressed in different forms are lost, which could lead to missing content during the retrieval process. For example, to find "magmatic" content, the keyword-based searching method finds all the fragments that contain the string "magmatic". However, many sub-categories under the concept of "magmatic", such as basalt, dacite, and tuff, exist in the geological domain. Although these concepts also belong to the category of "magmatic" in the geological domain ontology, they do not contain the keyword "magmatic" in their expressions; thus, fragments that contain these concepts are not discovered.

Table 4 shows that the precision rate and recall rate were both promoted with the semantic retrieval method, especially the recall rate. An exception occurred for "Volcanic rocks": the precision of the retrieval decreased compared to that of the old method because some unrelated items were extended as a search item during retrieval, which negatively affected the search.

In addition, semantic-related technologies were introduced into the framework. The GSISSP could automatically discover spatial information by building a retrieval frame ontology and combining the SWRL rules and the workflow. If we ask a specific class of geological question, the GSISSP could build appropriate workflows according to the predefined SWRL rules, which implies that the system could discover explicit content on the target theme or area and mine implicit information by customizing the related spatial analysis. Unstructured content discovery was added to our retrieval framework according to the work by Jung, Sun and Yuan [33]. Spatial content that is related to the target theme and area could be discovered and displayed on a map through our system. Similarly, unstructured content (for example, text, images, or tables) that is related to the target theme and area would be discovered, which would enable the user to fully exploit the information that is implied in the unstructured data. Therefore, the spatial results integrated the unstructured information to provide more comprehensive geological information.

7. Future work

In future studies, we plan to introduce data-mining algorithms and machine learning technologies based on available geological data contents. We will use data-mining algorithms to investigate deeper mining works and further enrich the associations in the geological domain ontology to discover potential knowledge and information from geological big data.

Acknowledgments: This project was supported by the National Science Foundation of China (Grant No. 41671400) and the Key Research Project, which is supported by the Natural Science Foundation of Hubei Province (Grant No. 2015CFA012). The authors thank the Development and Research Center of the China Geological Survey for providing technical support. We thank the NERCGIS (the National Engineering Research Center for Geographic Information System of China) for providing hardware support.

Author Contributions: Conceived and designed the experiments: Liang Wu, Lei Xue, Chaoling Li and Zhong Xie; Performed the experiments: Liang Wu, Lei Xue, Zhanlong Chen and Baode Jiang; Analyzed the data: Liang Wu, Chaoling Li, Zhong Xie and Xia Lv; Contributed reagents/materials/analysis tools: Lei Xue, Zhanlong Chen and Mingqiang Guo; Wrote the paper: Liang Wu, Lei Xue, Chaoling Li and Zhong Xie.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. O'Driscoll, A.; Daugelaite, J.; Sleator, R.D. "Big data", hadoop and cloud computing in genomics. *J. Biomed. Inform.* **2013**, *46*, 774–781. [CrossRef] [PubMed]
2. Evangelidis, K.; Ntouros, K.; Makridis, S.; Papatheodorou, C. Geospatial services in the cloud. *Comput. Geosci.* **2014**, *63*, 116–122. [CrossRef]
3. Sharma, S. Expanded cloud plumes hiding big data ecosystem. *Future Gener. Comput. Syst.* **2016**, *59*, 63–92. [CrossRef]
4. Yang, C.; Yu, M.; Hu, F.; Jiang, Y.; Li, Y. Utilizing cloud computing to address big geospatial data challenges. *Comput. Environ. Urban Syst.* **2017**, *61*, 120–128. [CrossRef]
5. Linh Manh, P.; El-Rheddane, A.; Donsez, D.; de Palma, N. Cirus: An elastic cloud-based framework for ubilytics. *Ann. Telecommun.* **2016**, *71*, 133–140.
6. Vera-Baquero, A.; Colomo-Palacios, R.; Molloy, O. Real-time business activity monitoring and analysis of process performance on big-data domains. *Telemat. Inform.* **2016**, *33*, 793–807. [CrossRef]
7. Wylot, M.; Cudre-Mauroux, P. Diplocloud: Efficient and scalable management of rdf data in the cloud. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 659–674. [CrossRef]
8. Xia, J.; Yang, C.; Liu, K.; Li, Z.; Sun, M.; Yu, M. Forming a global monitoring mechanism and a spatiotemporal performance model for geospatial services. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 375–396. [CrossRef]
9. Giachetta, R. A framework for processing large scale geospatial and remote sensing data in mapreduce environment. *Comput Graph.* **2015**, *49*, 37–46. [CrossRef]
10. Oweis, N.E.; Owais, S.S.; George, W.; Suliman, M.G.; Snasel, V. A survey on big data, mining: (tools, techniques, applications and notable uses). In *Intelligent Data Analysis and Applications*; Abraham, A., Jiang, X.H., Snasel, V., Pan, J.S., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 370, pp. 109–119.
11. Lomotey, R.K.; Deters, R. Towards knowledge discovery in big data. In Proceedings of the IEEE 8th International Symposium on Service Oriented System Engineering, Oxford, UK, 7–11 April 2014; IEEE: Oxford, UK, 2014.; pp. 181–191.
12. Kim, G.-H.; Trimi, S.; Chung, J.-H. Big-data applications in the government sector. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
13. Yang, C.; Huang, Q.; Li, Z.; Liu, K.; Hu, F. Big data and cloud computing: Innovation opportunities and challenges. *Int. J. Digit. Earth* **2017**, *10*, 13–53. [CrossRef]
14. Bhogal, J.; Choksi, I. Handling big data using NoSQL. In Proceedings of the 29th IEEE International Conference on Advanced Information Networking and Applications Workshops, Gwangju, Korea, 24–27 March 2015; IEEE: Gwangju, Korea, 2015.; pp. 393–398.
15. Lomotey, R.K.; Deters, R. Terms mining in document-based NoSQL: Response to unstructured data. In Proceedings of the 3rd IEEE International Congress on Big Data, BigData Congress, Anchorage, AK, USA, 27 June–2 July 2014; IEEE: Anchorage, AK, USA, 2014.; pp. 661–668.
16. Mazurek, M. Applying nosql databases for operationalizing clinical data mining models. In *Beyond Databases, Architectures and Structures*; Kozielski, S., Mrozek, D., Kasprowski, P., MalysiakMrozek, B., Kostrzewa, D., Eds.; Springer: Berlin, Germany, 2014; Volume 424, pp. 527–536.
17. Lomotey, R.K.; Deters, R. Unstructured data extraction in distributed NoSQL. In Proceedings of the 7th IEEE International Conference on Digital Ecosystems and Technologies: Smart Planet and Cyber Physical Systems as Embodiment of Digital Ecosystems, Menlo Park, CA, USA, 24–26 July 2013; IEEE: Menlo Park, CA, USA, 2013.; pp. 160–165.
18. Lomotey, R.K.; Deters, R. Topics and terms mining in unstructured data stores. In Proceedings of the IEEE 16th International Conference on Computational Science and Engineering, Sydney, Australia, 3–5 December 2013; Chen, J., Cuzzocrea, A., Yang, L.T., Eds.; IEEE: Sydney, Australia, 2013.; pp. 854–861.

19. Lomotey, R.K.; Deters, R. Real-time effective framework for unstructured data mining. In Proceedings of the 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Melbourne, Australia, 16–18 July 2013; pp. 1081–1088.
20. Witayangkurn, A.; Horanont, T.; Shibasaki, R. The design of large scale data management for spatial analysis on mobile phone dataset. *Asian J. Geoinform.* **2013**, *13*, 3.
21. Zhao, J.; Wang, L.; Tao, J.; Chen, J.; Sun, W.; Ranjan, R.; Kołodziej, J.; Streit, A.; Georgakopoulos, D. A security framework in g-hadoop for big data computing across distributed cloud data centres. *J. Comput. Syst. Sci.* **2014**, *80*, 994–1007. [CrossRef]
22. Zhong, Y.; Han, J.; Zhang, T.; Li, Z.; Fang, J.; Chen, G. Towards parallel spatial query processing for big spatial data. In Proceedings of the 2012 IEEE 26th International on Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), Shanghai, China, 21–25 May 2012; pp. 2085–2094.
23. Aji, A.; Wang, F.; Vo, H.; Lee, R.; Liu, Q.; Zhang, X.; Saltz, J. Hadoop gis: A high performance spatial data warehousing system over mapreduce. *Proc. VLDB Endow.* **2013**, *6*, 1009–1020. [CrossRef]
24. Eldawy, A.; Mokbel, M.F. A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. *Proc. VLDB Endow.* **2013**, *6*, 1230–1233. [CrossRef]
25. Zou, Z.Q.; Wang, Y.; Cao, K.; Qu, T.S.; Wang, Z.M. Semantic overlay network for large-scale spatial information indexing. *Comput. Geosci.* **2013**, *57*, 208–217. [CrossRef]
26. Verma, V.K.; Ranjan, M.; Mishra, P. Text mining and information professionals role, issues and challenges. In Proceedings of the 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (Ettlis), Noida, India, 6–8 January 2015; pp. 133–137.
27. Sirsat, S.R.; Chavan, V.; Deshpande, S.P. Mining knowledge from text repositories using information extraction: A review. *Sadhana* **2014**, *39*, 53–62. [CrossRef]
28. Abu, A.; Susan, L.L.H.; Sidhu, A.S.; Dhillon, S.K. Semantic representation of monogenean haptoral bar image annotation. *BMC Bioinform.* **2013**, *14*, 48. [CrossRef] [PubMed]
29. Kuo, C.L.; Hong, J.H. Interoperable cross-domain semantic and geospatial framework for automatic change detection. *Comput. Geosci.* **2016**, *86*, 109–119. [CrossRef]
30. Stock, K.; Stojanovic, T.; Reitsma, F.; Ou, Y.; Bishr, M.; Ortmann, J.; Robertson, A. To ontologise or not to ontologise: An information model for a geospatial knowledge infrastructure. *Comput. Geosci.* **2012**, *45*, 98–108. [CrossRef]
31. Cruz, S.A.B.; Monteiro, A.M.V.; Santos, R. Automated geospatial web services composition based on geodata quality requirements. *Comput. Geosci.* **2012**, *47*, 60–74. [CrossRef]
32. Li, W.; Yang, C.; Nebert, D.; Raskin, R.; Houser, P.; Wu, H.; Li, Z. Semantic-based web service discovery and chaining for building an arctic spatial data infrastructure. *Comput. Geosci.* **2011**, *37*, 1752–1762. [CrossRef]
33. Jung, C.-T.; Sun, C.-H.; Yuan, M. An ontology-enabled framework for a geospatial problem-solving environment. *Comput. Environ. Urban Syst.* **2013**, *38*, 45–57. [CrossRef]
34. Xiao, C.; Chen, N.; Wang, X.; Chen, Z. A semantic registry method using sensor metadata ontology to manage heterogeneous sensor information in the geospatial sensor web. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 63. [CrossRef]
35. Hu, Y.; Janowicz, K.; Prasad, S.; Gao, S. Enabling semantic search and knowledge discovery for arcgis online: A linked-data-driven. In Proceedings of the 18th AGILE International Conference on Geographic Information Science, AGILE 2015, Lisbon, Portugal, 9–12 June 2015; Kluwer Academic Publishers: Lisbon, Portugal, 2015.; pp. 107–124.
36. Ganesan, V.; Waheeta, H.S.; Srimathi, H. Jena with sparql to find indian natural plants used as medicine for diseases. In Proceedings of the International Conference on Internet Computing and Information Communications, Chennai, India, 12–14 February 2012; Sathiakumar, S., Awasthi, L.K., Masillamani, M.R., Sridhar, S.S., Eds.; Springer: Berlin, Germany, 2014.; pp. 225–237.
37. Alves, M.B.; Damasio, C.V.; Correia, N. Sparql commands in jena rules. In Proceedings of the 6th International Conference Knowledge Engineering and Semantic Web, KESW 2015, Moscow, Russia, 30 September–2 October 2015; Klinov, P., Mouromtsev, D., Eds.; Springer: Moscow, Russia, 2015.; pp. 253–262.
38. Thangsupachai, N.; Niwattanakul, S.; Chamnongsri, N. Learning object metadata mapping for linked open data. In *Emergence of Digital Libraries—Research and Practices*; Tuamsuk, K., Jatowt, A., Rasmussen, E., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 8839, pp. 122–129.

39. Chebotko, A.; Lu, S.; Fei, X.; Fotouhi, F. Rdfprov: A relational rdf store for querying and managing scientific workflow provenance. *Data Knowl. Eng.* **2010**, *69*, 836–865. [CrossRef]
40. Giordano, D.; Maiorana, F. Learning about the semantic web in an information systems oriented curriculum: A case study. In *Computer Supported Education*; Zvacek, S., Restivo, M.T., Uhomoibhi, J., Helfert, M., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 510, pp. 242–257.
41. Jang, B.; Ha, Y.-G. Transitivity reasoning for rdf ontology with iterative mapreduce. In Proceedings of the Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Taichung, Taiwan, 3–5 July 2013; IEEE: Taichung, Taiwan, 2013.; pp. 232–237.
42. Liu, S.; Zhang, F.; Zhai, Z. Modeling and discovering data services over sparql services. In Proceedings of the IEEE World Congress on Services, Anchorage, AK, USA, 27 June–2 July 2014; Zhang, L.J., Bahsoon, R., Eds.; IEEE: Anchorage, AK, USA, 2014.; pp. 169–173.
43. Jing, Y.; Jeong, D.; Baik, D.-K. Sparql graph pattern rewriting for owl-dl inference queries. *Knowl. Inf. Syst.* **2009**, *20*, 243–262. [CrossRef]
44. Song Wan, L.; Ni Li, X. Semantic query and reasoning system based on domain ontology. In Proceedings of the 2015 International Symposium on Computers & Informatics, Beijing, China, 17–18 January 2015; Liang, H., Wang, W., Eds.; Atlantis Press: Amsterdam, The Netherlands, 2015.; pp. 2524–2531.
45. Christodoulou, G.; Petrakis, E.G.M.; Batsakis, S. Qualitative spatial reasoning using topological and directional information in owl. In Proceedings of the IEEE 24th International Conference on Tools with Artificial Intelligence, Athens, Greece, 7–9 November 2012; IEEE: Athens, Greece, 2012.; pp. 596–602.
46. Herrero-Zazo, M.; Segura-Bedmar, I.; Hastings, J.; Martinez, P. Dinto: Using owl ontologies and swrl rules to infer drug-drug interactions and their mechanisms. *J. Chem. Inf. Model.* **2015**, *55*, 1698–1707. [CrossRef] [PubMed]
47. Orlando, J.P.; Musen, M.A.; Moreira, D.A. User extensible system to identify problems in owl ontologies and swrl rules. In *Rule Technologies: Foundations, Tools, and Applications*; Bassiliades, N., Gottlob, G., Sadri, F., Paschke, A., Roman, D., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9202, pp. 112–126.
48. Cantone, D.; Longo, C.; Nicolosi-Asmundo, M.; Santamaria, D.F. Web ontology representation and reasoning via fragments of set theory. In *Web Reasoning and Rule Systems*; TenCate, B., Mileo, A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9209, pp. 61–76.
49. Wu, L.; Xue, L.; Li, C.; Lv, X.; Chen, Z.; Guo, M.; Xie, Z. A geospatial information grid framework for geological survey. *PLoS ONE* **2015**, *10*, e0145312. [CrossRef] [PubMed]
50. White, T. Meet hadoop. In *Hadoop—The Definitive Guide*; Tsinghua University Press: Beijing, China, 2010; pp. 1–11.
51. White, T. The hadoop distributed filesystem. In *Hadoop—The Definitive Guide*; Tsinghua University Press: Beijing, China, 2010; pp. 44–79.
52. White, T. Zookeeper. In *Hadoop—The Definitive Guide*; Tsinghua University Press: Beijing, China, 2010; pp. 394–430.
53. George, L. Introduction. In *Hbase—The Definitive Guide*; POST & TELECOM PRESS: Beijing, China, 2013; pp. 1–26.
54. George, L. Advanced usage. In *Hbase—The Definitive Guide*; POST & TELECOM PRESS: Beijing, China, 2013; pp. 339–365.
55. Gruber, T.R. A translation approach to portable ontology specifications. *Knowl. Acquis.* **1993**, *5*, 199–220. [CrossRef]
56. Neches, R.; Fikes, R.E.; Finin, T.; Gruber, T.; Patil, R.; Senator, T.; Swartout, W.R. Enabling technology for knowledge sharing. *AI Mag.* **1991**, *12*, 36.
57. Giaretta, P.; Guarino, N. Ontologies and knowledge bases towards a terminological clarification. In *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*; Mars, N., Ed.; IOS Press: Amsterdam, The Netherlands, 1995; Volume 25, p. 32.
58. An, Y.; Zhao, B. *Geo Ontology Design and Comparison in Geographic Information Integration*; IEEE Computer Society: Washington, DC, USA, 2007; pp. 608–612.
59. Zhong, J.; Aydina, A.; McGuinness, D.L. Ontology of fractures. *J. Struct. Geol.* **2009**, *31*, 251–259. [CrossRef]

60. Li, C.; Song, M.; Lv, X.; Luo, X.; Li, J. The spatial data sharing mechanisms of geological survey information grid in p2p mixed network systems network architecture model. In Proceedings of the 2010 9th International Conference on Grid and Cooperative Computing (GCC), Nanjing, China, 1–5 November 2010; pp. 258–263.
61. Li, C. Geological domain ontology and its application. In *China Geological Survey Information Grid—Technology & Methodology*; Geological Publishing House: Beijing, China, 2013; pp. 27–51.
62. Li, C. The technical infrastructure of geological survey information grid. In Proceedings of the 2010 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010.