Hindawi
Scientific Programming
Volume 2017, Article ID 5072427, 13 pages
https://doi.org/10.1155/2017/5072427

Research Article

Intelligent Learning for Knowledge Graph towards Geological Data

Yueqin Zhu,[1,2] Wenwen Zhou,[2,3,4] Yang Xu,[2,3,4] Ji Liu,[2,3,4] and Yongjie Tan[1,2]

[1] *Development and Research Center, China Geological Survey, Beijing 100037, China*
[2] *Key Laboratory of Geological Information Technology, Ministry of Land and Resources, Beijing 100037, China*
[3] *School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing 100083, China*
[4] *Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China*

Correspondence should be addressed to Yueqin Zhu; yueqin_zhu@126.com and Yang Xu; b20160304@xs.ustb.edu.cn

Received 14 September 2016; Revised 21 December 2016; Accepted 12 January 2017; Published 16 February 2017

Academic Editor: HuaPing Liu

Knowledge graph (KG) as a popular semantic network has been widely used. It provides an effective way to describe semantic entities and their relationships by extending ontology in the entity level. This article focuses on the application of KG in the traditional geological field and proposes a novel method to construct KG. On the basis of natural language processing (NLP) and data mining (DM) algorithms, we analyze those key technologies for designing a KG towards geological data, including geological knowledge extraction and semantic association. Through this typical geological ontology extracting on a large number of geological documents and open linked data, the semantic interconnection is achieved, KG framework for geological data is designed, application system of KG towards geological data is constructed, and dynamic updating of the geological information is completed accordingly. Specifically, unsupervised intelligent learning method using linked open data is incorporated into the geological document preprocessing, which generates a geological domain vocabulary ultimately. Furthermore, some application cases in the KG system are provided to show the effectiveness and efficiency of our proposed intelligent learning approach for KG.

1. Introduction

Geological data is a variety of data and information accumulated in the geological research work and practical activities. Generally, the types of geological data are in a wide variety, including geological documents, geological books, geological information and journals, physical specimens, and electronic file data [1–3]. Due to the technical reasons, traditional storage modes may lead to the operations inefficient in queries, statistics, and updates; then they are not conducive to the application of checking, querying, and mining, which means the low ability of data services.

With the increasing economic and society, in the field of geological survey, geological data sharing service has become an important tool to measure the level of social and business management, which is significant in ensuring the sustainable development of geological work. The features of geological data include increasing volume, complex type, and long response time. Aiming at the geological application problems, the intelligent analysis and deep mining of geological data could reduce the repetitive working and the risk of geological survey [4, 5].

In recent years, knowledge service based on the knowledge graph (KG) technology and the search technology of semantic web has become a research hot spot in information service. In this case, the KG arises at the historic moment [6–10]. Drawing KG and conducting intelligent search based on KG have formed a mature methodology. For example, in Chinese, Sogou Knowledge Cube is the first KG introduced into the domestic search engine [11], which makes a reoptimization of search results through the integration of massive Internet fragmented information, and presents the core of the information to users. Baidu Zhixin is a new generation of Baidu search engine technology based on its KG [12]. There are four steps in the process of constructing the KG, including the named entity mining, attribute-value pair (AVP) mining, upper and lower relation mining, and related entity mining. Although there are some successful applications, it still has room for developing KG, and the applications still should be further strengthen, especially for the geological data.

In this article, the KG construction technology is applied in geology to implement intelligent analysis and deep mining of geological data. Through an unsupervised knowledge learning method for open data sources, we not only achieve self-learning process for a set of documents, but also form a geology glossary and complete the construction of KG. Through the research along this topic, promoting the geological materials information and social services has important value for the realization of intelligent geological survey.

The contributions of this article are as follows:

(1) *On the basis of linked open data and ontology learning strategy, we achieve the unsupervised learning and geological knowledge extraction for geological documents*. Through the processing steps of word segmentation, web crawler, keywords extraction, and relation extraction, the processing of geological documents and the deep mining of geological information are implemented.

(2) *Through the use of geological data sample, the geological ontology library, including entities, geology dictionary, and semantic link are obtained*. Firstly, we analyze the features of the geological data and acquire the geological ontology based on the geological knowledge extraction. Meanwhile, considering the specificity of the geological data, we design the corresponding entities combining with the geological dictionary and other specific documents. Secondly, through the processing of documents and web crawler using online encyclopedia, an expanded dictionary of geology and complete interconnection of semantic relations are proposed.

(3) *The KG towards geological data is proposed and the application system under Browser/Server (B/S) schema is also achieved*. Through the optimization of semantic relations and the storage of our knowledge base, we develop a framework for the application of the KG towards geological data, on the basis of the self-processing and self-expanding technologies towards geological documents. Combined with HTML5, JSP, Servlet, JDBC, and other advanced technologies, a B/S-based application system of KG is designed to realize the documents importing and processing, the intermediate results presenting, and expert intervening.

The rest of this article is organized as follows. Section 2 provides an analysis for background regarding the features of geological data and KG. The details of the intelligent learning scheme for KG, including framework, key technologies, and algorithms, are proposed in Section 3. Furthermore, the evaluation for our developed KG is conducted on two application experiments in Section 4. Finally, a conclusion is provided in Section 5.

2. Backgrounds

2.1. The Features of Geological Data. In recent years, with the increasing demand of geological data in production units and social masses, the geological data services are facing the dual demands of “digitization” and “socialization.” It is necessary to improve the contents and methods of geological data services, promoting the government departments to adapt to the development of the situation and achieve the transformation from the archival results to the service product [13]. Due to the features of reusability, reprocessing, and long-term service, the data information in mining industry, which has been accumulated over the years, could be called “big data.” Generally, geological data is mainly composed of structured and unstructured diversity data generated from geological actions. Its features are summarized as “5V,” that is, volume, variety, value, velocity, and veracity.

The requirement of maintaining the type and quantity of geological data grows with the long-term accumulation for data. It includes various types of electronic file data, such as documents, maps, database (map database, spatial database, and attribute database), pictures, charts, videos, audio, which might be structured, semistructured, and unstructured. Due to the technical reasons, this storage mode makes data query, statistics, updates, and other operations to the data not only inefficient, but also detrimental to the application, such as check, query, and mining, which leads to the low capability to the data service. Hence, it is significant to exploring how to apply the concept and technology of big data to organize massive geological data in the field of geology effectively and achieve the corresponding services [14].

Generally, the diversified fragmentation of complex geological unstructured data is one of the most striking features. There are mainly three contents that reflect to the data analysis and mining processing, including the establishment of content index library, search, and clustering recommendation [15]. Although it has achieved some results on this aspect [16, 17], with the development of the intelligent geological survey, multicategory whose content extension organization and search application are based on geological domain ontology will be an important direction of geological data repository construction in the future [18]. In view of this aspect, we can try to use the semantic link technology based on KG to eliminate the ambiguity of search, which could make the search engine use the search based on entity instead of character string. In addition, the Internet could also be used to provide rich resources for the KG, in order to realize the semantic link of big data, intelligent analysis, and mining of the geological large data accurately and effectively.

2.2. Knowledge Graph (KG). KG is also known as science knowledge graph, knowledge domain visualization, and knowledge domains map. It is a series of various graphs that show the development process of scientific knowledge and the structure relation [19]. It could describe the knowledge resource and its supporter, excavation, analysis, and construction and draw and display the knowledge and the mutual connection between them by using the visualization technology [6, 20, 21]. KG is a research method combining the theory and method of applied mathematics, graphics, information visualization technology, information science, and other disciplines with metrology citation analysis, cooccurrence analysis, and other methods to show the core framework, the history of development, the frontier domain, and overall knowledge structure of the discipline through visual graph. It shows the dynamic development of knowledge and the complex domain knowledge through data mining, information processing, knowledge measurement, and graphics rendering.

Most works regarding the KG originated from Google KG. It is essentially a semantic network. The nodes represent entities or concepts and the edges represent a variety of semantic relations between entities and concepts. Moreover, the motivation of KG is from a series of practical applications, including semantic search, machine answering, information retrieval, electronic reading, and online learning. Now, some companies, such as Baidu, Sogou, have launched their own KG.

Our researchers have developed many applications around KG, while illustrating different perspectives in their process. For example, in the process of visual analysis of Chinese science literature, it showed the time sequence distribution, journal distribution, and author distribution of scientific literature during the past 30 years [22]. According to cocitation analysis for the authors of the quotation in 24 kinds of information science core journals, the KG of information science was drawn [23]. Based on the example of education in China, KG was drawn to evaluate excellent scientific research institutions through word frequency analysis, high frequency author statistics, high yield author cooperative network, and other methods [24].

In addition to the above applications, many scholars have also carried out some works in KG. Hook showed that KG has four purposes (i.e., discovery, understanding, communication, and education) and six aspects of application (i.e., microcosmic display of specific areas, macroscopic visualization of subject, assisting in the education course teaching, saving document knowledge in coordination, facilitating the use of digital library, and displaying knowledge dissemination) [25]. It indicated that KG could be used in displaying the overall structure of the domain knowledge, analyzing retrieval result visually, grasping the overall knowledge of discipline and evolution situation of visualization knowledge, and grasping the rapid variation of the knowledge [26]. Meanwhile, information fusion as a key issue plays an important role in developing a KG. To deal with this issue, a novel approach was proposed for exemplar extraction through structured sparse, considering not only the reconstruction capability and the sparsity, but also the diversity and robustness [27]. Furthermore, based on previous work, a joint kernel sparse coding model was developed to solve the multifinger tactile sequence classification problem, where all of the coding vectors were encouraged to share the same sparsity support pattern [28].

KG applications increase rapidly in recent years, which cover some disciplines of natural science and social science, and show the osmotic tendency towards other disciplines. Drawing KG and mining KG have formed a high mature methodology. However, the function of KG has not been fully applied, and the application still needs to be further strengthened. So far, only little attention has been paid to the geological data field. Hence, it is necessary and important to consider these particular objects.

3. Intelligent Learning for Knowledge Graph

3.1. The System Framework. The construction of KG towards geological data consists of two logical components: knowledge extraction and knowledge management. The former mainly learns the corresponding geological knowledge through unsupervised processing and including five steps, which are word segmentation, frequency statistics, web crawler, keywords extraction, and relation extraction. The latter is basically composed of two parts: knowledge graph storage and retrieval. The specific processes are shown in Figure 1.

3.2. Knowledge Extraction. Knowledge extraction is a key step in the construction of knowledge graph, as well as in the processing of geological documents. Knowledge extraction in this article, through an unsupervised knowledge learning method based on an open source, and the geological domain vocabulary and knowledge graph would be formed through the automatic learning of a large number of geological documents. The flow of knowledge extraction is shown in Figure 2.

Knowledge extraction has three major steps, including data sources analysis, entity/concept extraction, and relation extraction.

3.2.1. The Analysis of the Available Data Sources

(1) Text. Texts are the most abundant data source. It is difficult to learn knowledge from texts due to their nonstructural property. In this article, we obtain a large number of geological professional texts from library.

(2) Internet Encyclopedia. Internet encyclopedias (e.g., Wikipedia, Baidu baike, and Baike.com) are the large-scale free encyclopedias that allow users to edit almost any article accessible. Through technical tools such as web crawler, we obtain knowledge from Internet encyclopedias continuously, which could be updated and expanded automatically.

Although the contents of encyclopedias exist with the form of web pages, there are still a lot of structured information. Since all of encyclopedias have their own classification system, category labels are used to organize a large number of entries. In general, each entry has category label, which could be used to label its own type. In addition, most of entries have multiple labels. For example, the category labels of “Steve Jobs” could be “20th-century American business people,” “American billionaires,” “American computer business people,” and many others in Wikipedia.

This article mainly focuses on Chinese information in Internet encyclopedias. Wikipedia is considered the Internet’s largest and most popular general reference book. However, Chinese content in Wikipedia is not perfect. On the one

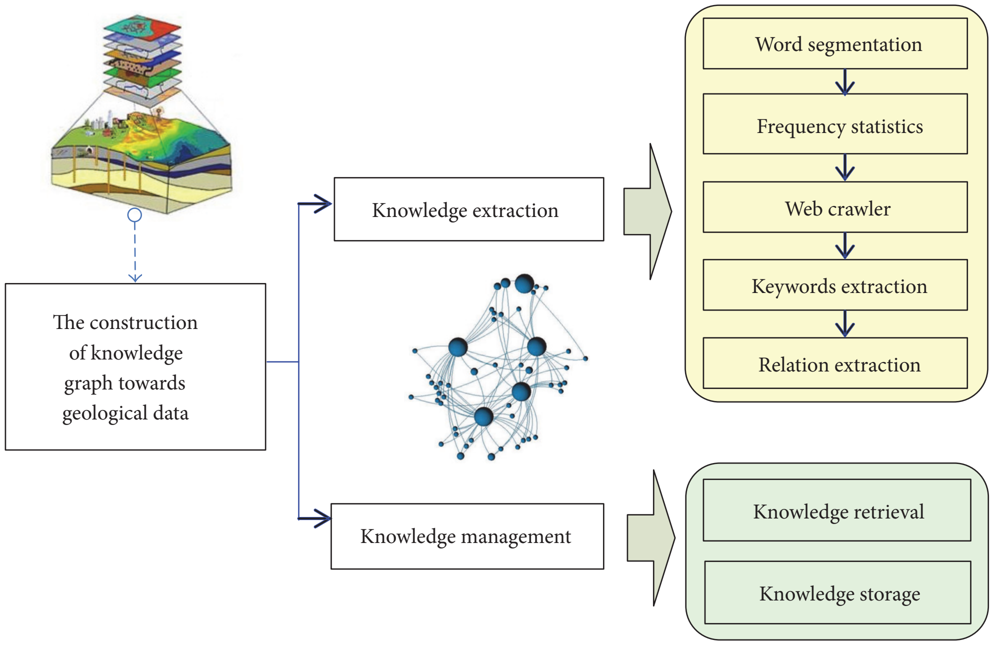


Figure 1: The logical structure of knowledge graph construction towards geological data.

hand, the total number of entries is insufficient. And, the contents of the articles in Wikipedia are also relatively short, and some parts of them are translated from other languages directly, which are lacking the expression exactly in Chinese. Consequently, we make use of Baike.com instead of Wikipedia as the data source of web crawler in this article.

3.2.2. Entity/Concept Extraction. Entity/concept extraction mainly starts from these two data sources. We could filter out entities or concepts of geology directly by combining the information after text processing with category labels of Baike.com. Therefore, the entity/concept extraction includes four bottom-up steps: word segmentation, frequency statistics, web crawler, and keywords extraction.

The technology of HanLP could be used in the word segmentation, stop word filtration, and frequency statistics. Motivated by the TextRank algorithm, word segmentation used in this article is as follows. First of all, we use HanLP standard tokenizer to process documents, which are divided into different parts of speech words. Secondly, the custom data dictionary and extended stop list are designed. Finally, we filter out the word with little relevance to retrieving content and only retain the designated part of speech through the method of TextRank algorithm. Meanwhile, we also filter out the stop words, so as to achieve the effect of keyword extraction.

In terms of web crawler, we mainly consider to crawl the category labels of entries in the Internet encyclopedia by an automated tool Selenium, which could open the HtmlUnit browser, search entries, and access to class label information via programming by custom. Specifically, the method for online encyclopedia crawler is as follows. When we want to get information about a word "*n*," we should open our browser first. Then, we search and open encyclopedia interface of "*n*." We can locate and save category label elements by XPath finally.

In terms of the keywords extraction, according to the geological dictionary and the category labels, we could exactly determine whether the words in the segmentation results belong to the geological keywords or not. Through the statistical characteristics of Wikipedia category labels, we extract some keywords, including geography, mining, marine, rock, hydrology, environment, natural disasters, biology, city, air, oil, roads, plants, energy, metallurgy, and civil. We put all crawled category labels into a map collection. By calling the containsKey method of map, we can determine whether the collected object contains the keywords, if the answer is yes, this object is defined as a geological entity.

3.2.3. Relation Extraction. The purpose of relation extraction is nontaxonomic relation extraction of the association rule analysis in data mining and the Internet encyclopedia. The correlation between two geological terminologies is acquired by association rule analysis. And the category relationship of terminologies is acquired through crawling Internet encyclopedia.

The basic principle of association rule is that if the two concepts or entities frequently appeared in the same unit (e.g., a document, a paragraph, or a sentence), we could make sure there exist some relationships between them. We do not care about the specific semantic relations between two concepts, but the correlated degree between them. Hence, judging the correlated degree between two concepts through cooccurrence analysis in a document is more important. With the increase of the number of documents processed, there would be a higher correlated degree if the two concepts frequently appeared together. This method is also motivated by the process of human reading and learning. However, this method is just suitable to be

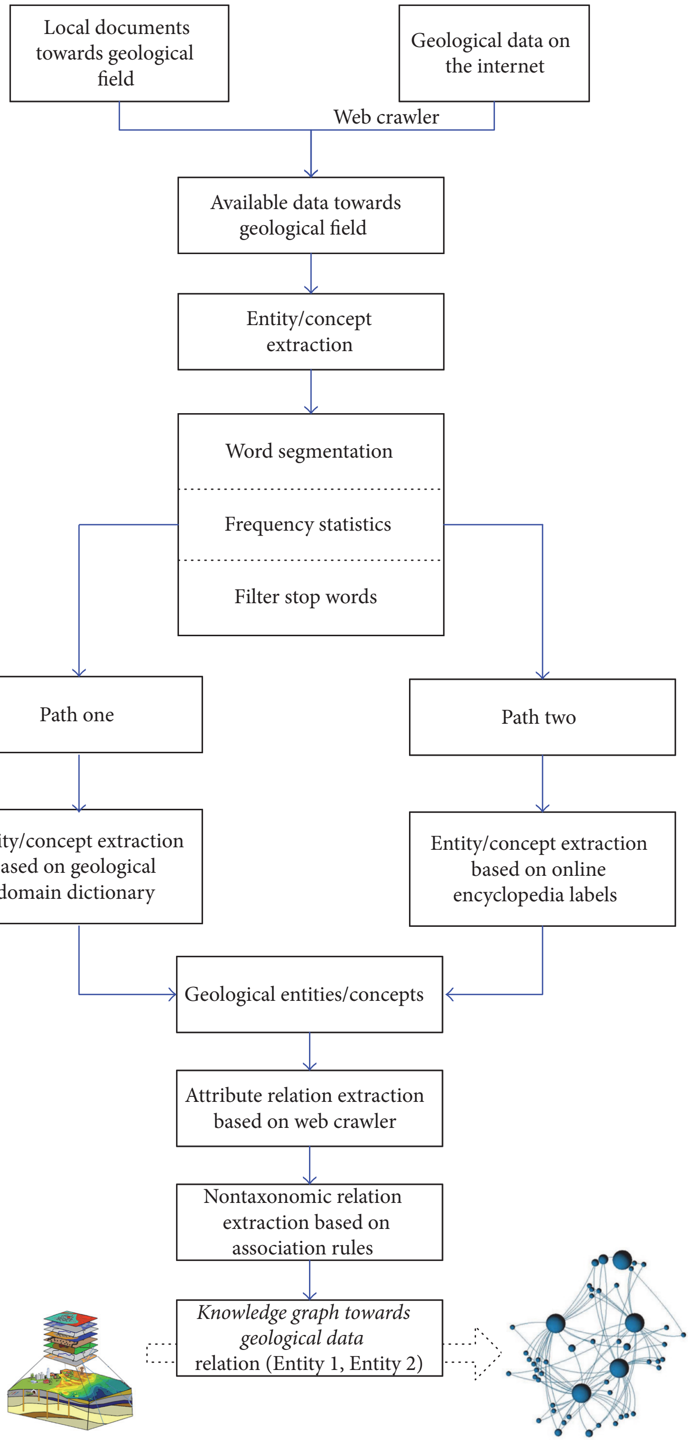

Local documents towards geological field
Geological data on the internet
Web crawler
Available data towards geological field
Entity/concept extraction
Word segmentation
Frequency statistics
Filter stop words
Path one
Path two
Entity/concept extraction based on geological domain dictionary
Entity/concept extraction based on online encyclopedia labels
Geological entities/concepts
Attribute relation extraction based on web crawler
Nontaxonomic relation extraction based on association rules
Knowledge graph towards geological data
relation (Entity 1, Entity 2)

Figure 2: The flow of knowledge extraction.

employed for dealing with large number of documents; when the number of documents is small, this method would be inefficient.

Meanwhile, the purpose of crawling Internet encyclopedia is to obtain relationships between concepts and entities by making use of the open data source in the online encyclopedia. As mentioned above, here we mainly consider the category relationships.

Using the above two methods, the rule of our relation extraction is as follows. In terms of correlated degree, we set a relational degree R for each concept, where the initial value of R is 0. After processing a document, the correlation between all the words which appeared in the document is increased by 1. The value of R updates once in the process of dealing with document each time. Furthermore, each concept has category labels as their property.

Table 1: The attributes of tables in background database.

Table name	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
“articles”	ID	Content	Date	Path	—
“words”	ID	Content	Frequency	Label	—
“re_words_articles”	ID	ID1	ID2	Frequency	—
“re_words_words”	ID	ID1	ID2	ID3	Count
“dictionary”	Name	Label	—	—	—

3.3. Knowledge Management. Knowledge management considers how to show the knowledge acquired through above steps in a visualized way. The main technical methods are the database storage and retrieval.

3.3.1. Database Storage. Considering the actual needs of the geological field, the system uses MySQL database as the background database. MySQL database is one of the best relational database management systems in web application, which has a small size, fast storage and retrieval speed, and low cost.

In our system, the entities and relationships acquired by processing geological documents are stored in a special database. Through JDBC technology, the background database operations, such as CRUD, are allowed. There are five tables in our database. Table “articles” stores the information about the documents processed, including ID, name, added time, and local storage path of documents. Table “words” stores the information about the words filtered out from the results of segmentation, including ID, content, frequency, and category labels of words. Table “re_words_words” stores the correlation information between two geological terms.

The attributes of those tables in our background database are in Table 1.

3.3.2. Knowledge Graph Retrieval. Retrieval can be carried out by users only after storing the knowledge extracted from documents in our database. Based on B/S working schema, the browser makes a post request to the back-end server after users input the search words. Meanwhile, back-end server responds to the request, getting the words submitted and the number of nodes that need to be rendered (it is set to 20 as a default value). The retrieved words are set to the key nodes and retrieved in our database. Then, it returns the results to the browser. The returned contents include ID, content and category labels of node, and ID of correlated documents.

3.3.3. Backstage Management System of KG. Backstage management system of KG is designed to facilitate the process of documents and the operation of database for users, mainly including login page, geological documents processing page, and expert intervention page.

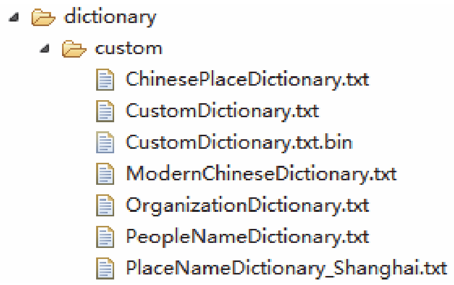
Two login modes could be chosen when users enter the login page by inputting the URL in the browser. Users could enter the geological document processing page if logged in as an administrator. Users also could enter the page of expert intervention if logged in as an expert. The browser submits the form, including name, password, and login mode. Subsequently, the users authorization would be checked by server, and users could enter the relevant page after verified.

On the page of geological documents processing, users could input the document name and storage path. And background module gets the form data submitted by users when the button “submit” is clicked on. The background module enters the stage of document processing and the results are stored in background database if all of this input data is valid. On the page of expert intervention, the experts have the right to add and delete the correlation between two words. For example, when adding a correlation, the experts enter the two words in the input box and click the button “submit.” The browser submits these two words to the background module, and the background module judges whether there is a correlation between them or not. If the association does not exist, the background module would add a correlation, which is defined as “expert-defined.”

3.4. The Key Algorithms. The prototype system of KG towards geological big data is designed and accordingly implemented using B/S architecture and HTTP protocol, which includes natural language processing (NLP), data mining, web application development, and other related technologies. Key technologies and solutions involved during the process of system development are described as follows.

3.4.1. The Automatic Chinese Segmentation Technology: HanLP. HanLP is a Java toolkit composed of a series of models and algorithms, whose target is to promote the application of NLP in the production environment. HanLP supports Chinese word segmentation. Its functions include N shortest path word segmentation, CRF word segmentation, index word segmentation, and user defined dictionary. Specifically, they are named entity recognition, keyword extraction, phrase extraction, Pinyin conversion, conversion between simplified and complex, and dependency parsing (i.e., MaxEnt dependency parsing, CRF dependency parsing). The characteristics of HanLP are perfect function, efficient performance, clear architecture, new corpus, and being customizable.

(1) TextRank Algorithm. Making the use of TextRank for Chinese word segmentation mainly includes word segmentation, delete stop words, and iterative voting. The basic idea of TextRank Chinese word segmentation is as follows: dividing the original text into sentences first, filtering out the withdrawal in each sentence, and only retaining the specified part of speech word. From it, we could get a set of sentences and a set of words. Then each word would be as a node



```
dictionary
  custom
    ChinesePlaceDictionary.txt
    CustomDictionary.txt
    CustomDictionary.txt.bin
    ModernChineseDictionary.txt
    OrganizationDictionary.txt
    PeopleNameDictionary.txt
    PlaceNameDictionary_Shanghai.txt
```

Figure 3: User defined dictionary.

in the TextRank by using the method of matrix iterative convergence [29]. The window size is set to K, and we assume that a sentence is constructed by the following words:

$$w_1, w_2, w_3, w_4, w_5, \ldots, w_n, \quad (1)$$

where $w_1, w_2, \ldots, w_k, w_2, w_3, \ldots, w_{k+1}, w_3, w_4, \ldots, w_{k+2}$ are all in a window. There is an undirected and unweighted edge between any two words corresponding to the node in a window.

With the above composition diagram, we could calculate the weight of each word node. Then, the iterative formula in TextRank algorithm is as follows:

$$\begin{aligned} \mathrm{WS}\left(V_i\right) = {} & (1-d) + d \\ & \times \sum_{V_j \in \mathrm{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \mathrm{Out}(V_j)} w_{jk}} \mathrm{WS}\left(V_j\right), \end{aligned} \quad (2)$$

where $d \in [0, 1]$ is a damping factor, V_i is a given node, $\mathrm{In}(V_i)$ are the set of nodes that point to it (predecessors), $\mathrm{Out}(V_i)$ is the set of nodes that node V_i points to (successors), and $\mathrm{WS}(V_i)$ is the weight of node V_i.

(2) User Defined Dictionary. HanLP word segmentation supports the function of custom dictionary, where our custom dictionary designed is shown in Figure 3.

We add a large number of words that could help the word segmentation of geological documents in the custom dictionary effectively. Here, the "CustomDictionary" includes 21,742 geological words, the "OrganizationDictionary" includes 31,926 institutional nouns, the "ChinesePlaceDictionary" includes 90,558 place names, the "PeopleNameDictionary" includes 50,192 personal names, and the "ModernChineseDictionary" includes 207,964 modern Chinese additional words. Among them, "CustomDictionary" is a dictionary defined by a global user which could add, delete, and affect all word segmentation at any time.

3.4.2. Internet Encyclopedia Crawler Based on Selenium.

On the basis of our analysis mentioned above, it is effective and efficient to integrate the online encyclopedia crawler technology into the processing flow of geological documents, which need to get the category labels of words obtained by word segmentation in Wikipedia. The mainstream method of crawler is implemented using URL address webs which can be obtained through depth or breadth first search strategy. Here, the web site that we need to crawl is fixed (i.e., http://www.baike.com/), and we already have the target word (i.e., word segmentation results). Then, it is available to introduce the automated crawler technology. Here, we use Selenium automated crawler.

Selenium automation test browser is mainly applied to the automated testing of the web application, while supporting all management task automation based on web. By embedding the Selenium IDE plug-in into the browser, the recording and playback functions of a simple browser operation could be achieved.

It should be noted that Selenium provides a highly rapid and convenient way for the fixed web crawler. Here, we use Selenium to control HtmlUnit, a virtual browser that Java comes with, which serves the purpose of automated crawler. The specific process mainly includes opening the HtmlUnit browser, reading a search word "*n*," retrieving by opening the encyclopedia interface of the retrieved word "*n*," getting category labels according to category label elements by XPath, and finally closing the browser.

Implementation details of Internet encyclopedia crawler are as follows:

(1) Open HtmlUnit browser: *static final WebDriver driver = new HtmlUnitDriver()*

(2) Open the interface of search word "*n*": *driver.get ("http://www.baike.com/wiki/"+n)*

(3) Locate the label element:

List<WebElement> elements = driver.findElements(By.xpath("//dL[@id='show_tag']/dd/a"))

3.4.3. Java Web Development Based on Servlet and Java Server Pages (JSP).

Java Servlet is a Java program that extends the capabilities of a server. Although Servlets could respond to any types of requests, they implement applications hosted on web servers usually. Such Web Servlets are the Java counterpart to other dynamic web content technologies, such as PHP and ASP.NET.

Servlets are often used to process and store a Java class in Java EE that conforms to the Java Servlet API, a standard for implementing Java classes which could respond to the requests. And, Servlets could communicate over any clientCserver protocol, but they are often used with the HTTP protocol. So, "Servlet" is often used as shorthand for "HTTP Servlet." Thus, a software developer should use a Servlet to add dynamic content to a web server by using the Java platform. The generated content is HTML but may be other data such as XML. Servlets could maintain state in session variables across many server transactions by using HTTP cookies or rewriting URLs.

Servlets could be generated from JSP by the JavaServer pages compiler automatically. Architecturally, JSP could be viewed as a high-level abstraction of Java Servlets. It allows Java code and certain predefined actions to be interleaved with static web markup content, such as HTML, with the resulting page being compiled and executed on the server to deliver a document. JSP are translated into Servlets at

Table 2: The Servlets and their key functions.

Servlet name	Key functions
“Myservlet.java”	It is used in the retrieval of KG, and it gets the form data submitted by the user and retrieves them.
“Myservlet2.java”	It is used in the second retrieval. When clicking on some word in the page, the user can get the graph of this word.
“LoginServlet.java”	It is used in the login function of the backstage management system of KG, and it gets the form data submitted by the user and enters the response page.
“AddServlet.java”	It is used while adding a relationship in expert intervention page.
“DelServlet.java”	It is used while deleting a relationship in expert intervention page.
“CoreServlet.java”	It is used while showing the intermediate processing for geological documents.

runtime, and each JSP Servlet is cached and reused until the original JSP is modified.

Servlets can complete the following tasks:

(1) The web container initializes the Servlet instance; then the Servlet instance could read data that has been provided in the HTTP request.

(2) The Servlet instance could create and return a dynamic response page to the client.

(3) The Servlet instance could access server resources, such as files and database.

(4) The Servlet instance could prepare dynamic data for the JSP and create a response page with JSP.

In this article, the Servlets and their key functions that we design under com.servlet package are shown in Table 2.

In summary, the software platforms and development environments in our system are as follows. Operating system is Windows 7. Programming language is Java. Programming environment is MyEclipse 10. Web development environment is Tomcat + Severlet + JSP. Web crawler environment is Selenium + HtmlUnit.

4. Experiments and Evaluation

4.1. Processing for a Single Document. Processing for a single geological document is as shown in Figure 4. We can see that user who logs as the administrator enters the document processing page. Then, the user inputs the name and storage path of document and clicks on the submit button. A background module gets the form data submitted by the administrator and determines if this document exists in the local path. The background module enters the stage of document processing when all of this input data are valid. The document is converted into a long string. Then the background module would cut word segmentation by HanLP, filter out stop word, and select out the geological terminologies. The result after intermediate processing is showed in the document processing page.

The document is processed using the similar method in [30].

4.1.1. The Result Analysis of Segmentation

(1) Some results of segmentation in our KG system are shown in Figure 5. After translating it from Chinese into English, the updated version of Figure 5 is shown in Figure 6.

(2) Some results of segmentation by NLPIR systems of Beijing Institute of Technology [31], which is a popular NLP system, are shown in Figure 7. After translating it from Chinese into English, the updated version of Figure 7 is shown in Figure 8.

According to the process in [30], some results of segmentation in our KG system are showed in Figure 5. Meanwhile, Figure 6 shows some results of segmentation in NLPIR systems. By comparing these two figures, we can find that the results processed in our system are more valuable and satisfactory. For example, to these geological terminologies, such as “华北陆块 (North China craton),” “高于庄组 (Gaoyuzhuang Formation),” “下马岭组 (Xiamaling Formation),” “铁岭组 (Tieling Formation),” and “吕梁运动 (Luliang Movement),” our KG system can accurately cut word segmentation. However, in NLPIR systems, many geological terminologies are cut inaccurately.

4.1.2. Word Frequency Statistics. The results of word frequency statistics are as shown in Figure 9. After translating it from Chinese into English, the updated version of Figure 9 is shown in Figure 10. We can see that our system can count word frequency correctly for the result set of segmentation, such as “杨庄组 (Yangzhuang Formation)/13,” “下马岭组 (Xiamaling Formation)/8,” “长石石英砂岩 (Feldspathic Quartz Sandstone)/1,” and “同位素年龄 (Isotope Age)/1.”

4.1.3. Keywords Extraction. Figure 11 shows the results of keywords extraction. We normally consider that the terminologies contained in the title and subtitles are basically the keywords for the document. Therefore, the result of keywords extraction in Figure 11 includes the critical position “华北陆块 (North China Craton),” three key stratigraphic units “高于庄组 (Gaoyuzhuang Formation),” “杨庄组 (Yangzhuang Formation),” and “下马岭组 (Xiamaling Formation),” the key stratigraphic unit “元古界 (Proterozoic Erathem),” and the major stratigraphic relationship “不整合面 (Unconformities).” In summary, our keywords extraction has satisfying results.

4.1.4. Internet Encyclopedia Crawler. The results of category labels crawled from the Internet encyclopedia (http://www.baike.com/) are as shown in Figure 12, including geological terminologies of segmentation result sets and the category labels. After translating it from Chinese into English, the updated version of Figure 12 is shown in Figure 13.

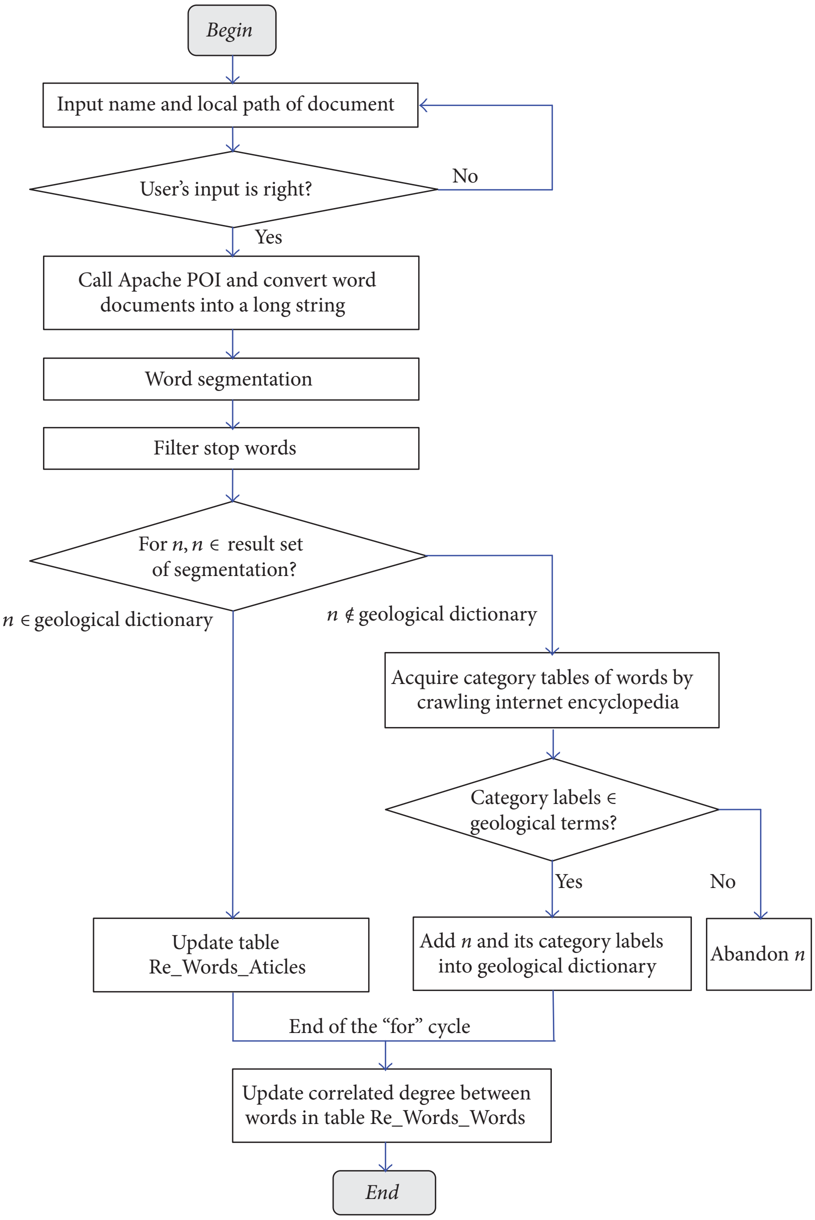

Begin
Input name and local path of document
User's input is right?
No
Yes
Call Apache POI and convert word documents into a long string
Word segmentation
Filter stop words
For n, $n \in$ result set of segmentation?
$n \in$ geological dictionary
$n \notin$ geological dictionary
Acquire category tables of words by crawling internet encyclopedia
Category labels $\in$ geological terms?
Yes
No
Update table Re_Words_Aticles
Add n and its category labels into geological dictionary
Abandon n
End of the "for" cycle
Update correlated degree between words in table Re_Words_Words
End

FIGURE 4: Processing for geological documents.

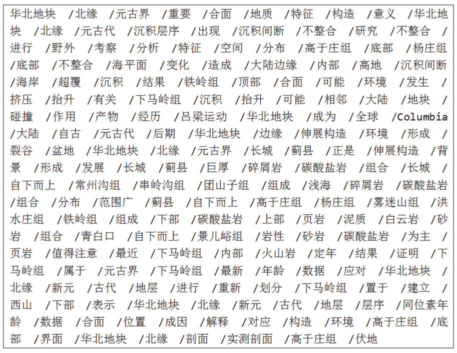

华北地块 /北缘 /元古界 /重要 /合面 /地质 /特征 /构造 /意义 /华北地
块 /北缘 /元古代 /沉积层序 /出现 /沉积间断 /不整合 /研究 /不整合 /
进行 /野外 /考察 /分析 /特征 /空间 /分布 /高于庄组 /底部 /杨庄组
/底部 /不整合 /海平面 /变化 /造成 /大陆边缘 /内部 /高地 /沉积间断
/海岸 /超覆 /沉积 /结果 /铁岭组 /顶部 /合面 /可能 /环境 /发生 /
挤压 /抬升 /有关 /下马岭组 /沉积 /抬升 /可能 /相邻 /大陆 /地块 /
碰撞 /作用 /产物 /经历 /吕梁运动 /华北地块 /成为 /全球 /Columbia
/大陆 /自古 /元古代 /后期 /华北地块 /边缘 /伸展构造 /环境 /形成 /
裂谷 /盆地 /华北地块 /北缘 /元古界 /长城 /蓟县 /正是 /伸展构造 /背
景 /形成 /发展 /长城 /蓟县 /巨厚 /碎屑岩 /碳酸盐岩 /组合 /长城 /
自下而上 /常州沟组 /串岭沟组 /团山子组 /组成 /浅海 /碎屑岩 /碳酸盐岩
/组合 /分布 /范围广 /蓟县 /自下而上 /高于庄组 /杨庄组 /雾迷山组 /洪
水庄组 /铁岭组 /组成 /下部 /碳酸盐岩 /上部 /页岩 /泥质 /白云岩 /砂
岩 /组合 /青白口 /自下而上 /景儿峪组 /岩性 /砂岩 /碳酸盐岩 /为主 /
页岩 /值得注意 /最近 /下马岭组 /内部 /火山岩 /定年 /结果 /证明 /下
马岭组 /属于 /元古界 /下马岭组 /最新 /年龄 /数据 /应对 /华北地块 /
北缘 /新元 /古代 /地层 /进行 /重新 /划分 /下马岭组 /置于 /建立 /
西山 /下部 /表示 /华北地块 /北缘 /新元 /古代 /地层 /层序 /同位素年
龄 /数据 /合面 /位置 /成因 /解释 /对应 /构造 /环境 /高于庄组 /底
部 /界面 /华北地块 /北缘 /剖面 /实测剖面 /高于庄组 /伏地

FIGURE 5: Some results of segmentation.

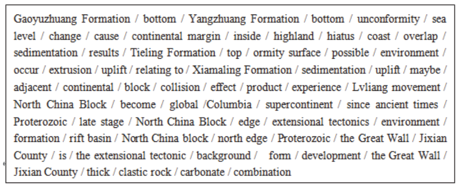
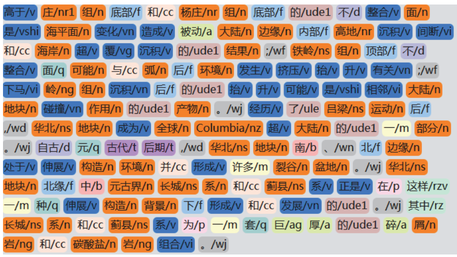

Gaoyuzhuang Formation / bottom / Yangzhuang Formation / bottom / unconformity / sea
level / change / cause / continental margin / inside / highland / hiatus / coast / overlap /
sedimentation / results / Tieling Formation / top / ormity surface / possible / environment /
occur / extrusion / uplift / relating to / Xiamaling Formation / sedimentation / uplift / maybe /
adjacent / continental / block / collision / effect / product / experience / Lvliang movement /
North China Block / become / global /Columbia / supercontinent / since ancient times /
Proterozoic / late stage / North China Block / edge / extensional tectonics / environment /
formation / rift basin / North China block / north edge / Proterozoic / the Great Wall / Jixian
County / is / the extensional tectonic / background / form / development / the Great Wall /
Jixian County / thick / clastic rock / carbonate / combination

FIGURE 6: The updated version of Figure 5 after translating it from Chinese into English.

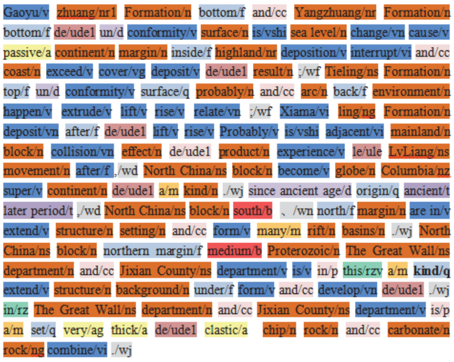
4.2. *Searching in KG.*

The specific process of retrieval in KG is shown in Figure 14. We can see from this figure that the first step users need to do is to input the retrieved word



```
高于/v 庄/nr1 组/n 底部/f 和/cc 杨庄/nr 组/n 底部/f 的/ude1 不/d 整合/v 面/n
是/vshi 海平面/n 变化/vn 造成/v 被动/a 大陆/n 边缘/n 内部/f 高地/nr 沉积/v 间断/vi
和/cc 海岸/n 超/v 覆/vg 沉积/v 的/ude1 结果/n ;/wf 铁岭/ns 组/n 顶部/f 不/d
整合/v 面/q 可能/n 与/cc 弧/n 后/f 环境/n 发生/v 挤压/v 抬/v 升/v 有关/vn ;/wf
下马/vi 岭/ng 组/n 沉积/vn 后/f 的/ude1 抬/v 升/v 可能/v 是/vshi 相邻/vi 大陆/n
地块/n 碰撞/vn 作用/n 的/ude1 产物/n 。/wj 经历/v 了/ule 吕梁/ns 运动/n 后/f
,/wd 华北/ns 地块/n 成为/v 全球/n Columbia/nz 超/v 大陆/n 的/ude1 一/m 部分/n
。/wj 自古/d 元/q 古代/t 后期/t ,/wd 华北/ns 地块/n 南/b 、/wn 北/f 边缘/n
处于/v 伸展/v 构造/n 环境/n 并/cc 形成/v 许多/m 裂谷/n 盆地/n 。/wj 华北/ns
地块/n 北缘/f 中/b 元古界/n 长城/ns 系/n 和/cc 蓟县/ns 系/v 正是/v 在/p 这样/rzv
一/m 种/q 伸展/v 构造/n 背景/n 下/f 形成/v 和/cc 发展/vn 的/ude1 。/wj 其中/rz
长城/ns 系/n 和/cc 蓟县/ns 系/v 为/p 一/m 套/q 巨/ag 厚/a 的/ude1 碎/a 屑/n
岩/ng 和/cc 碳酸盐/n 岩/ng 组合/vi 。/wj
```

Figure 7: Some results of segmentation in NLPIR system.



```
Gaoyu/v zhuang/nr1 Formation/n bottom/f and/cc Yangzhuang/nr Formation/n
bottom/f de/ude1 un/d conformity/v surface/n is/vshi sea level/n change/vn cause/v
passive/a continent/n margin/n inside/f highland/nr deposition/v interrupt/vi and/cc
coast/n exceed/v cover/vg deposit/v de/ude1 result/n ;/wf Tieling/ns Formation/n
top/f un/d conformity/v surface/q probably/n and/cc arc/n back/f environment/n
happen/v extrude/v lift/v rise/v relate/vn ;/wf Xiama/vi ling/ng Formation/n
deposit/vn after/f de/ude1 lift/v rise/v Probably/v is/vshi adjacent/vi mainland/n
block/n collision/vn effect/n de/ude1 product/n experience/v le/ule LvLiang/ns
movement/n after/f ,/wd North China/ns block/n become/v globe/n Columbia/nz
super/v continent/n de/ude1 a/m kind/n ./wj since ancient age/d origin/q ancient/t
later period/t ,/wd North China/ns block/n south/b 、/wn north/f margin/n are in/v
extend/v structure/n setting/n and/cc form/v many/m rift/n basins/n ./wj North
China/ns block/n northern margin/f medium/b Proterozoic/n The Great Wall/ns
department/n and/cc Jixian County/ns department/v is/v in/p this/rzv a/m kind/q
extend/v structure/n background/n under/f form/v and/cc develop/vn de/ude1 ./wj
in/rz The Great Wall/ns department/n and/cc Jixian County/ns department/v is/p
a/m set/q very/ag thick/a de/ude1 clastic/a chip/n rock/n and/cc carbonate/n
rock/ng combine/vi ./wj
```

Figure 8: The updated version of Figure 7 after translating it from Chinese into English.

and click on the "search" button. Background module gets the form data submitted by the user and sets it as the key node. Furthermore, the background module retrieves in our database to filter out the terminologies that have a relational degree with the key node in top 20 (default) and shows them in a graph.

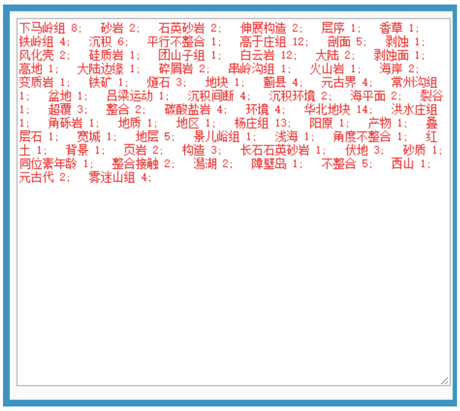
4.2.1. The Comparison of Different Retrieval Processing Stages

(1) After processing one geological document, the results of retrieving "变质岩 (metamorphic rock)" are in Figure 15.

(2) After processing 100 geological documents, the results of retrieving "变质岩 (metamorphic rock)" are in Figure 16.

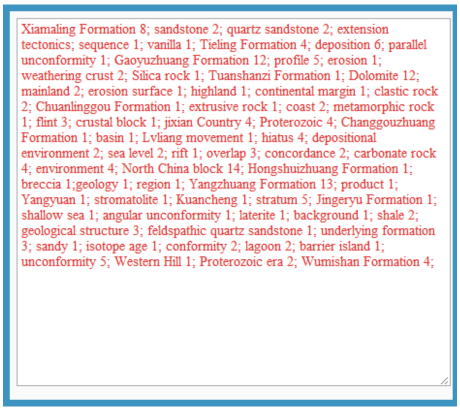
Figures 15 and 16 show the results of the KG retrieval. The orange node represents the retrieved word "变质岩 (metamorphic rock)." The blue nodes represent the terminologies, which have a top-20 relational degree with the orange node, such as "同位素年龄 (isotope age)" and "砂岩 (sandstone)." When the mouse is placed on some node, we could acquire its ID and category labels.

From the comparison of two retrieval processing stages, we could see that the results of KG have been improved with a growing number of documents processed. When the number of processed documents is 1, the retrieval results have little relevance with the retrieved word. However, when the number is 100, we could get entities that have a very close relationship with "变质岩 (metamorphic rock)," such as



```
下马岭组 8;   砂岩 2;   石英砂岩 2;   伸展构造 2;   层序 1;   香草 1;
铁岭组 4;   沉积 6;   平行不整合 1;   高于庄组 12;   剖面 5;   剥蚀 1;
风化壳 2;   硅质岩 1;   团山子组 1;   白云岩 12;   大陆 2;   剥蚀面 1;
高地 1;   大陆边缘 1;   碎屑岩 2;   串岭沟组 1;   火山岩 1;   海岸 2;
变质岩 1;   铁矿 1;   燧石 3;   地块 1;   蓟县 4;   元古界 4;   常州沟组
1;   盆地 1;   吕梁运动 1;   沉积间断 4;   沉积环境 2;   海平面 2;   裂谷
1;   超覆 3;   整合 2;   碳酸盐岩 4;   环境 4;   华北地块 14;   洪水庄组
1;   角砾岩 1;   地质 1;   地区 1;   杨庄组 13;   阳原 1;   产物 1;   叠
层石 1;   宽城 1;   地层 5;   景儿峪组 1;   浅海 1;   角度不整合 1;   红
土 1;   背景 1;   页岩 2;   构造 3;   长石石英砂岩 1;   伏地 3;   砂质 1;
同位素年龄 1;   整合接触 2;   潟湖 2;   障壁岛 1;   不整合 5;   西山 1;
元古代 2;   雾迷山组 4;
```

Figure 9: The results of word frequency statistics.



```
Xiamaling Formation 8; sandstone 2; quartz sandstone 2; extension
tectonics; sequence 1; vanilla 1; Tieling Formation 4; deposition 6; parallel
unconformity 1; Gaoyuzhuang Formation 12; profile 5; erosion 1;
weathering crust 2; Silica rock 1; Tuanshanzi Formation 1; Dolomite 12;
mainland 2; erosion surface 1; highland 1; continental margin 1; clastic rock
2; Chuanlinggou Formation 1; extrusive rock 1; coast 2; metamorphic rock
1; flint 3; crustal block 1; jixian Country 4; Proterozoic 4; Changgouzhuang
Formation 1; basin 1; Lvliang movement 1; hiatus 4; depositional
environment 2; sea level 2; rift 1; overlap 3; concordance 2; carbonate rock
4; environment 4; North China block 14; Hongshuizhuang Formation 1;
breccia 1;geology 1; region 1; Yangzhuang Formation 13; product 1;
Yangyuan 1; stromatolite 1; Kuancheng 1; stratum 5; Jingeryu Formation 1;
shallow sea 1; angular unconformity 1; laterite 1; background 1; shale 2;
geological structure 3; feldspathic quartz sandstone 1; underlying formation
3; sandy 1; isotope age 1; conformity 2; lagoon 2; barrier island 1;
unconformity 5; Western Hill 1; Proterozoic era 2; Wumishan Formation 4;
```

Figure 10: The updated version of Figure 9 after translating it from Chinese into English.

"花岗岩 (granite)," "岩浆 (magma)," and "火山岩 (volcanic rock)."

In addition, we could get the following information from the above results.

(1) The top 20 geological terminologies associated with "变质岩 (metamorphic rock)."

(2) The category labels for every geological terminology.

(3) The ID of documents in which both words appear.

4.2.2. Searching More Words. Furthermore, some complicated phrases and sentences can also be processed correctly. For example, when inputting "侵入岩和沉积岩 (intrusive rock and sedimentary rock)," the background module can cut word segmentation into two keywords "侵入岩 (intrusive rock)" and "沉积岩 (sedimentary rock)," retrieve them, and get the terminologies that have a relational degree with the key node in top 20. The results are as shown in Figure 17.

Similarly, we could get the following information from the above results.

(1) We can get the top 20 geological terminologies associated with "侵入岩 (intrusive rock)" and "沉积岩 (sedimentary rock)."

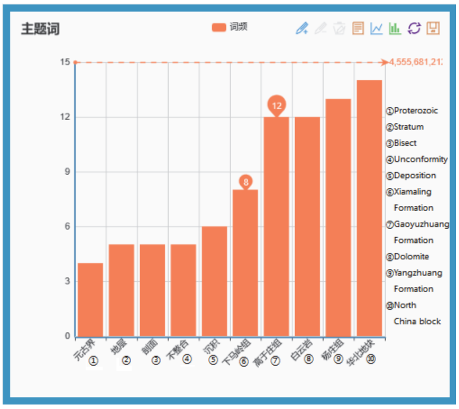
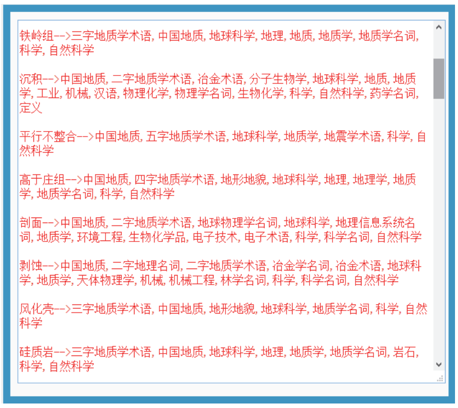


Figure 11: The results of keywords extraction.


铁岭组-->三字地质学术语, 中国地质, 地球科学, 地理, 地质, 地质学, 地质学名词, 科学, 自然科学

沉积-->中国地质, 二字地质学术语, 冶金术语, 分子生物学, 地球科学, 地质, 地质学, 工业, 机械, 汉语, 物理化学, 物理学名词, 生物化学, 科学, 自然科学, 药学名词, 定义

平行不整合-->中国地质, 五字地质学术语, 地球科学, 地质学, 地震学术语, 科学, 自然科学

高于庄组-->中国地质, 四字地质学术语, 地形地貌, 地球科学, 地理, 地理学, 地质学, 地质学名词, 科学, 自然科学

剖面-->中国地质, 二字地质学术语, 地球物理学名词, 地球科学, 地理信息系统名词, 地质学, 环境工程, 生物化学品, 电子技术, 电子术语, 科学, 科学名词, 自然科学

剥蚀-->中国地质, 二字地理名词, 二字地质学术语, 冶金学名词, 冶金术语, 地球科学, 地质学, 天体物理学, 机械, 机械工程, 林学名词, 科学, 科学名词, 自然科学

风化壳-->三字地质学术语, 中国地质, 地形地貌, 地球科学, 地质学名词, 科学, 自然科学

硅质岩-->三字地质学术语, 中国地质, 地球科学, 地理, 地质学, 地质学名词, 岩石, 科学, 自然科学

Figure 12: The results of Internet encyclopedia crawler.

(2) We can get the category labels for every geological terminology in KG.

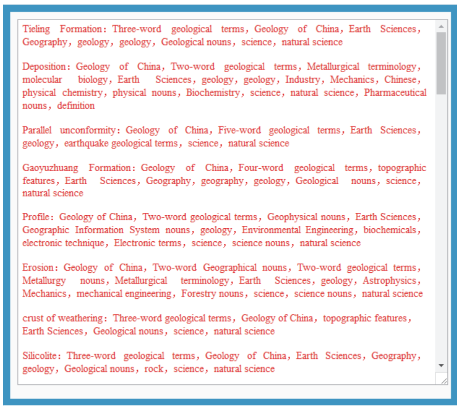
(3) We can get the ID of documents in which both words appear.

(4) While retrieving the two words, we can get the documents in which both words appear, and it achieves mining of implicit related documents.

(5) In addition, we can see the following:

(i) In terms of “侵入岩 (intrusive rock),” there exists a connecting line between “侵入岩 (intrusive rock)” and “花岗岩 (granite),” which means that there exists a high degree of correlation between them. However, there is no connecting line between “侵入岩 (intrusive rock)” and “泥岩 (mudstone),” which means that there exists a low correlation between them.

(ii) In terms of “沉积岩 (sedimentary rock),” there exists a connecting line between “沉积岩 (sedimentary rock)” and “泥岩 (mudstone).” However, there is no connecting line between “沉积岩 (sedimentary rock)” and “花岗岩 (granite).”


Tieling Formation：Three-word geological terms，Geology of China，Earth Sciences，Geography，geology，geology，Geological nouns，science，natural science

Deposition：Geology of China，Two-word geological terms，Metallurgical terminology，molecular biology，Earth Sciences，geology，geology，Industry，Mechanics，Chinese，physical chemistry，physical nouns，Biochemistry，science，natural science，Pharmaceutical nouns，definition

Parallel unconformity：Geology of China，Five-word geological terms，Earth Sciences，geology，earthquake geological terms，science，natural science

Gaoyuzhuang Formation：Geology of China，Four-word geological terms，topographic features，Earth Sciences，Geography，geography，geology，Geological nouns，science，natural science

Profile：Geology of China，Two-word geological terms，Geophysical nouns，Earth Sciences，Geographic Information System nouns，geology，Environmental Engineering，biochemicals，electronic technique，Electronic terms，science，science nouns，natural science

Erosion：Geology of China，Two-word Geographical nouns，Two-word geological terms，Metallurgy nouns，Metallurgical terminology，Earth Sciences，geology，Astrophysics，Mechanics，mechanical engineering，Forestry nouns，science，science nouns，natural science

crust of weathering：Three-word geological terms，Geology of China，topographic features，Earth Sciences，Geological nouns，science，natural science

Silicolite：Three-word geological terms，Geology of China，Earth Sciences，Geography，geology，Geological nouns，rock，science，natural science

Figure 13: The updated version of Figure 12 after translating it from Chinese into English.

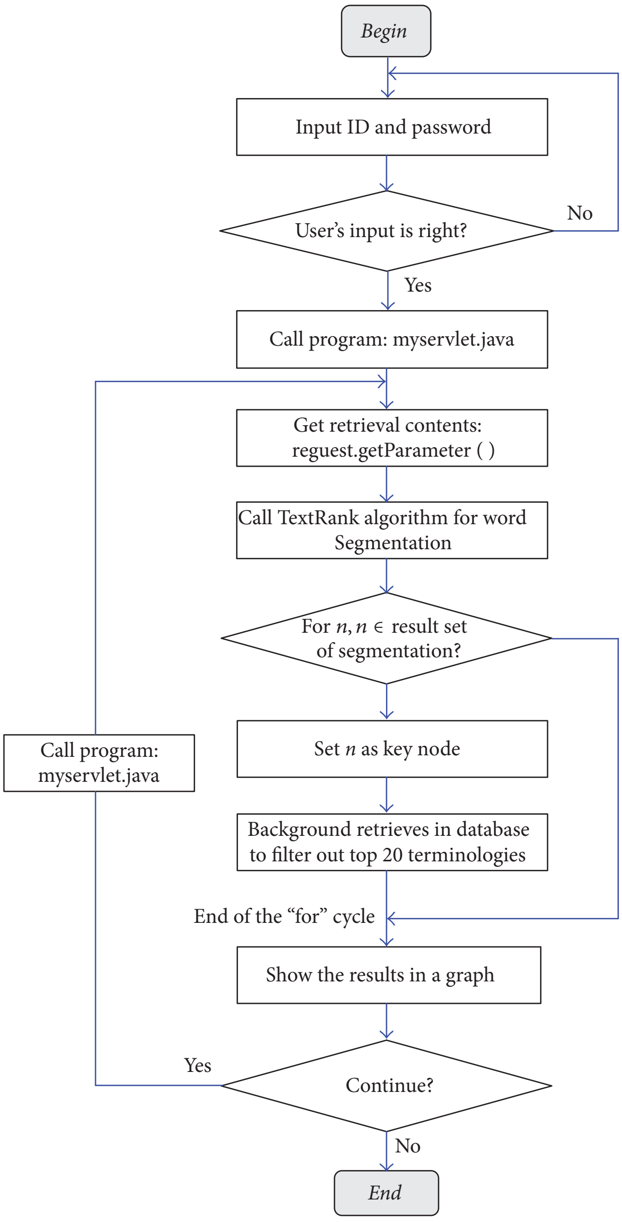

Begin
Input ID and password
User's input is right?
No
Yes
Call program: myservlet.java
Get retrieval contents: reguest.getParameter ( )
Call TextRank algorithm for word Segmentation
For n, $n \in$ result set of segmentation?
Set n as key node
Background retrieves in database to filter out top 20 terminologies
Call program: myservlet.java
End of the “for” cycle
Show the results in a graph
Yes
Continue?
No
End

Figure 14: The specific process of retrieval.

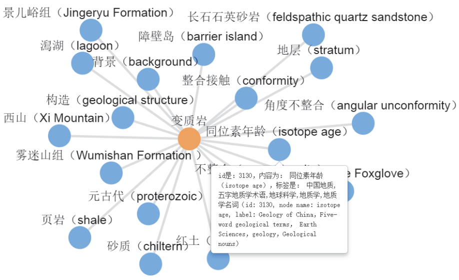

景儿峪组（Jingeryu Formation）
长石石英砂岩（feldspathic quartz sandstone）
障壁岛（barrier island）
潟湖（lagoon）
地层（stratum）
背景（background）
整合接触（conformity）
构造（geological structure）
角度不整合（angular unconformity）
西山（Xi Mountain）
变质岩
同位素年龄（isotope age）
雾迷山组（Wumishan Formation）
e Foxglove）
元古代（proterozoic）
页岩（shale）
砂质（chiltern）
红土
id是：3130，内容为： 同位素年龄（isotope age），标签是： 中国地质，五字地质学术语，地球科学，地质学，地质学名词（id: 3130, node name: isotope age, label: Geology of China, Five-word geological terms, Earth Sciences, geology, Geological nouns）

Figure 15: An experimental result after processing one geological document.

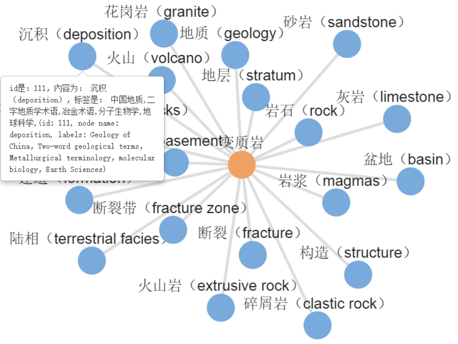


Figure 16: An experimental result after processing 100 geological documents.

Geological professionals know that "泥岩 (mudstone)" is a kind of "沉积岩 (sedimentary rock)" and "花岗岩 (granite)" is a kind of "侵入岩 (intrusive rock)." Hence, there is a high correlation between "侵入岩 (intrusive rock)" and "花岗岩 (granite)," as well as between "沉积岩 (sedimentary rock)" and "泥岩 (mudstone)." All the results we could see from the KG are our acquired learning information after processing 100 documents. Through this example, we could indicate that our KG system could provide valuable and accurate information in most cases. The more documents we process, the more accurate correlations we can get from the KG system.

4.2.3. The Formation of Geological Domain Dictionary. When processing geological documents, new geological terminologies and their category labels are obtained from web crawler. And they are added in our expanded geological domain dictionary.

In our experiments, the original number of words in geological domain dictionary is 11,062. And after processing 100 documents, the number of words in geological dictionary is 13,227. Some results of geological domain dictionary are in Figure 18, including geological terminologies and their corresponding category labels.

5. Conclusion

This article proposes a novel approach to constructing KG towards geological data. The proposed approach uses unsupervised learning method with linked open data to process geological documents and extract knowledge directly.

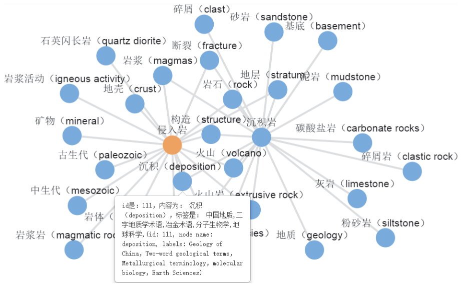
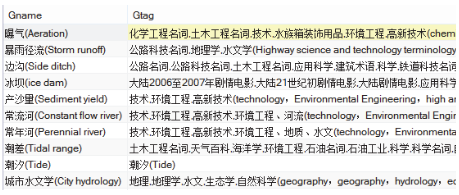


Figure 17: An experimental result after searching several key geological words.



Gname	Gtag
曝气(Aeration)	化学工程名词.土木工程名词.技术.水族箱装饰用品.环境工程.高新技术(chem
暴雨径流(Storm runoff)	公路科技名词.地理学.水文学(Highway science and technology terminology
边沟(Side ditch)	公路名词.公路科技名词.土木工程名词.应用科学.建筑术语.科学.铁道科技名词
冰坝(ice dam)	大陆2006至2007年剧情电影.大陆21世纪初剧情电影.大陆剧情电影.应用科学
产沙量(Sediment yield)	技术.环境工程.高新技术(technology，Environmental Engineering，high ar
常流河(Constant flow river)	技术.环境工程.高新技术.环境工程、河流(technology，Environmental Engir
常年河(Perennial river)	技术.环境工程.高新技术.环境工程、地质、水文(technology，Environmenta
潮差(Tidal range)	土木工程名词.天气百科.海洋学.环境工程.石油名词.石油工业.科学.科学名词.
潮汐(Tide)	潮汐(Tide)
城市水文学(City hydrology)	地理.地理学.水文.生态学.自然科学(geography，geography，hydrology，ec

Figure 18: The geological domain dictionary.

Through this approach, we accordingly achieve an effective self-learning process for documents, form a geology glossary, and complete the construction of KG based on the technologies of documents processing and dictionary expanding. Furthermore, we design an application system of KG on the basis of B/S working schema. Finally, the test on a large number of geological documents is conducted and some satisfactory results are obtained. In the future work, aiming at the features of geological data, the knowledge extracting approach in the KG is further improved to get more accurate entities and relations.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgments

This work was partially supported by the Special Funds Project for Scientific Research of Public Welfare Industry from the Ministry of Land and Resources of China under Grant 201511079 and the National Key Technologies R&D Program of China under Grant 2015BAK38B01.

References

[1] T. Zhang, Y. Du, T. Huang, and X. Li, "Stochastic simulation of geological data using isometric mapping and multiple-point geostatistics with data incorporation," *Journal of Applied Geophysics*, vol. 125, pp. 14–25, 2016.

[2] M. G. Runge, M. S. Bebbington, S. J. Cronin, J. M. Lindsay, and M. R. Moufti, "Integrating geological and geophysical data to improve probabilistic hazard forecasting of Arabian Shield volcanism," *Journal of Volcanology and Geothermal Research*, vol. 311, pp. 41–59, 2016.

[3] L. Zhang, "Improvement of K-means algorithm and its applications in analysis of geological exploration seismic data," *Electronic Journal of Geotechnical Engineering*, vol. 20, no. 12, pp. 4423–4434, 2015.

[4] Y. Zhu, Y. Tan, R. Li, and X. Luo, "Cyber-physical-social-thinking modeling and computing for geological information service system," in *Proceedings of the 4th International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI '15)*, Beijing, China, October 2015.

[5] X. Luo, D. Zhang, L. T. Yang, J. Liu, X. Chang, and H. Ning, "A kernel machine-based secure data sensing and fusion scheme in wireless sensor networks for the cyber-physical systems," *Future Generation Computer Systems*, vol. 61, pp. 85–96, 2016.

[6] D. Le-Phuoc, H. Nguyen Mau Quoc, H. Ngo Quoc, T. Tran Nhat, and M. Hauswirth, "The graph of things: a step towards the live knowledge graph of connected things," *Journal of Web Semantics*, vol. 37-38, pp. 25–35, 2016.

[7] D. Danciulescu and M. Colhon, "Systems of knowledge representation based on stratified graphs. Application in natural language generation," *Carpathian Journal of Mathematics*, vol. 32, no. 1, pp. 49–62, 2016.

[8] A. Ballatore, M. Bertolotto, and D. C. Wilson, "A structural-lexical measure of semantic similarity for geo-knowledge graphs," *ISPRS International Journal of Geo-Information*, vol. 4, no. 2, pp. 471–492, 2015.

[9] N. Jayaram, A. Khan, C. Li, X. Yan, and R. Elmasri, "Querying knowledge graphs by example entity tuples," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 10, pp. 2797–2811, 2015.

[10] B. Kamsu-Foguem and D. Noyes, "Graph-based reasoning in collaborative knowledge management for industrial maintenance," *Computers in Industry*, vol. 64, no. 8, pp. 998–1013, 2013.

[11] ZhiLiFang, http://baike.sogou.com/v66616234.htm.

[12] BaiduZhiXin, http://yingxiao.baidu.com/product/site/zhixin/.

[13] M. M. Song, Z. Li, B. Zhou, and C. L. Li, "Cloud computing model for big geological data processing," *Applied Mechanics and Materials*, vol. 475-476, pp. 306–311, 2014.

[14] M. Cao and L. Lu, "Nonparametric test models of geological hybrid parents based on big data," *ICIC Express Letters*, vol. 9, no. 9, pp. 2491–2498, 2015.

[15] C. Li, J. Li, H. Zhang, A. Gong, and D. Wei, "Big data application architecture and key technologies of intelligent geological survey," *Geological Bulletin of China*, vol. 34, no. 7, pp. 1288–1299, 2015.

[16] P. Vermeesch and E. Garzanti, "Making geological sense of 'big data' in sedimentary provenance analysis," *Chemical Geology*, vol. 409, pp. 20–27, 2015.

[17] G. Yan, Q. Xue, K. Xiao, J. Chen, J. Miao, and H. Yu, "An analysis of major problems in geological survey big data," *Geological Bulletin of China*, vol. 34, no. 7, pp. 1273–1279, 2015.

[18] C. L. Wu, G. Liu, X. L. Zhang, Z. W. He, and Z. T. Zhang, "Discussion on geological science big data and its applications," *Chinese Science Bulletin*, vol. 61, no. 16, pp. 1797–1807, 2016.

[19] Q. Liu, Y. Li, H. Duan, Y. Liu, and Z. Qin, "Knowledge graph construction techniques," *Jisuanji Yanjiu yu Fazhan/Computer Research and Development*, vol. 53, no. 3, pp. 582–600, 2016.

[20] H. Liu, F. Sun, B. Fang, and X. Zhang, "Robotic room-level localization using multiple sets of sonar measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 1, pp. 2–13, 2017.

[21] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-tactile fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, no. 99, pp. 1–13, 2016.

[22] Z. Y. Liu and C. B. Zhang, "Review of the 30-year studies of the methodology of science and technology in China-Based on the bibliometric analysis of journal articles," *Studies in Philosophy of Science and Technology*, vol. 31, no. 4, pp. 82–89, 2014.

[23] Y. Zhao and Y. Z. Sha, "The Knowledge mapping analysis on the research of information science: based on ACA," *Library Tribune*, vol. 28, no. 6, pp. 63–69, 2008.

[24] J. M. Tang, "Descriptive research of excellent scientific organizations based on bibliometric-setting national educational courses for example," *Journal of Intelligence*, vol. 29, no. 4, pp. 5–9, 2010.

[25] P. A. Hook, "Domain maps: purposes, history, parallels with cartography, and applications," in *Proceedings of the 11th International Conference Information Visualization (IV '07)*, Zurich, Switzerland, July 2007.

[26] Y. Zhang, P. J. H. Hu, S. A. Brown, and H. Chen, "Knowledge mapping for rapidly evolving domains: a design science approach," *Decision Support Systems*, vol. 50, no. 2, pp. 415–427, 2011.

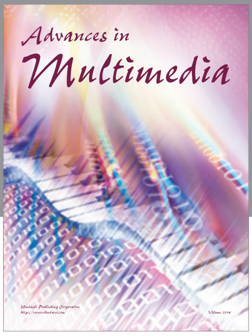
[27] H. Liu, Y. Liu, and F. Sun, "Robust exemplar extraction using structured sparse coding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 8, pp. 1816–1821, 2015.

[28] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: kernel sparse coding methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 656–665, 2016.

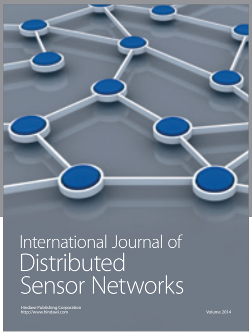
[29] A. Altman and M. Tennenholtz, "Ranking system: the PageRank axioms," in *Proceedings of the ACM Conference on Electronic Commerce*, pp. 1–8, 2005.

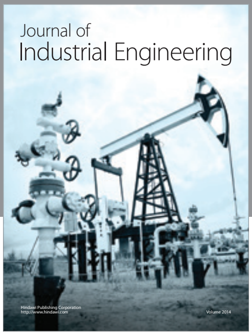
[30] Y. Q. Qu, Q. R. Meng, S. X. Ma, L. Li, and G. L. Wu, "Geological characteristics of unconformities in Mesoproterozoic successions in the northern margin of North China Block and their tectonic implications," *Earth Science Frontiers*, vol. 17, no. 4, pp. 112–127, 2010.

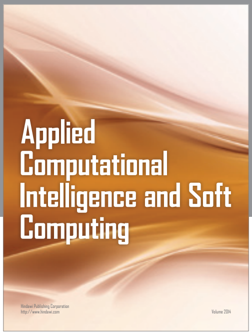
[31] Natural Language Processing and Information Retrieval (NLPIR) Sharing Platform, http://www.nlpir.org/.

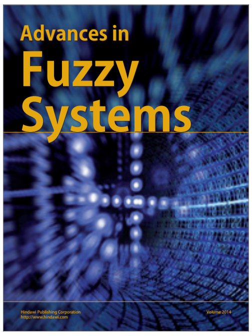

Advances in
Multimedia

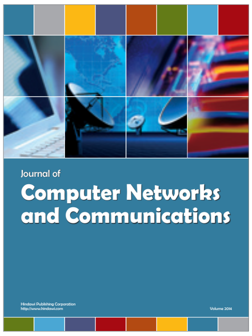
The Scientific
World Journal

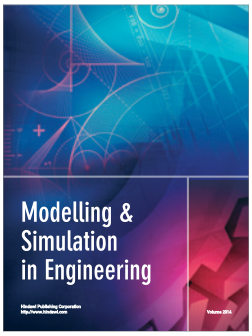

International Journal of
Distributed
Sensor Networks

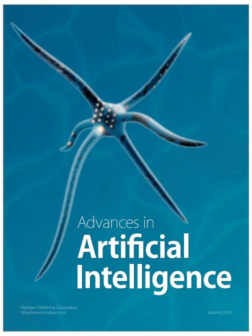

Journal of
Industrial Engineering

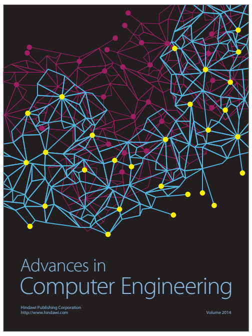

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

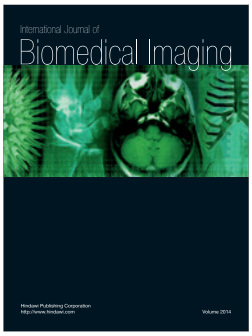

Journal of
Computer Networks
and Communications

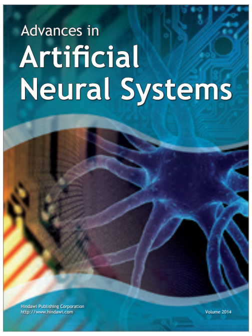

Modelling &
Simulation
in Engineering

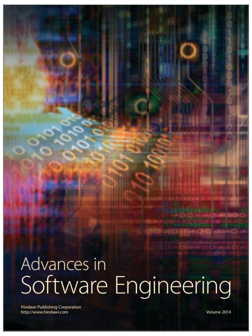

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

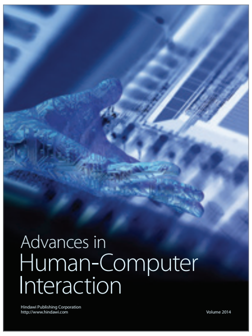
International Journal of
Computer Games
Technology

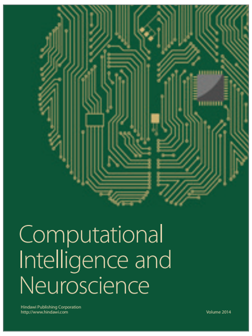

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering