Text Content Pushing Technology Research Based on Location and Topic

Dongqi Wei 1,2, Jianxin Wei 3,4, Naheman Wumuti 4 and Baode Jiang 2

1 Xi'an Center of Geological Survey, CGS, Xi'an, China
2 National Engineering Research Center for geographic Information System, China University of Geosciences(Wuhan), Wuhan, China
3 College of Resources and Environment Sciences, Xinjiang University, Urumqi, China
4 Xinjiang Land and Resources Information Center, Urumqi, China

Abstract. In the field, geological workers usually want to obtain related geological background information in the working area quickly and accurately. This information exists in the massive geological data, text data is described in natural language accounted for a large proportion. This paper studied location information extracting method in the mass text data; proposed a geographic location—geological content—geological content related algorithm based on Spark and Mapreduce2, finally classified content by using KNN, and built the content pushing system based on location and topic. It is running in the geological survey cloud, and we have gained a good effect in testing by using real geological data.

Keyword. text data; content pushing; big data; location; topic;

1. Introduction
The arrival of the big data era has brought a good opportunity to give full play to the value of geological data. Big data is a kind of high-level information productivity; it promotes the change of the information production mode and also promotes the development of geological data application mode, meets the practical needs of geological workers is the key to large data technology to landing in the field of geological survey. Field work is the characteristic of geological survey, how to perceive the location of the field geological workers and the related topics (such as formation, structure, geography situation, etc.) in the vicinity, to form an active and fast information content pushing capability, is a new direction of geological data applications, is also the embodiment of intelligent geology survey [1] and it has high research value and application prospects.

Long before the arrival of the big data era, industry using traditional data (here refers to the digital data) production methods and has accumulated abundant outcome data. Some of the information contained in these data has a high value for field workers, who work in the field, but this information is often collected in advance before out of the field, and it is hard to obtain real-time in the field. Therefore, it is the key point of this paper that how to sort out the information point of mass geological data reasonably and design the fast content pushing system for the field workers.

Finally, it should be noted that this paper uses the concept of "content" to represent a specific data and information, which is the minimum information processing unit of the system. Geological information
is the expression of geological information in a particular location space, it is clear that the geological content will be associated with some geological location. In addition, the geological content and geological location of the original geological data are recorded in the text, and the geological position is a relative position concept, which is difficult to be quantified (coordinate), therefore, how to establish the geographic location—geological location—geological content related method is the key problem in this paper.

2. Architecture
First of all, the data source of the system is a huge amount of unstructured geological data accumulated in the early stage, which is mainly presented in the form of geological survey report. To make the system has the ability of intelligent information pushing service for field workers, this requires the system consists of data service center, the wireless network (or communication satellites), satellite positioning system, mobile terminal. Its complete system structure and service mode is shown in figure 1.

System obtains the position information of the field geological workers, according to the theme of its subscription, through wireless communication network to push content information in the vicinity of its position. Data service center is the core part of the system with respect to other facilities. The data service center mainly provides content pushing services, location-based services and GIS services. These services are through a series of basic components to support top application. The system structure is shown in figure 2. Its important composition includes:

Location-based service (LBS): LBS obtains user's location coordinates, uses position encoder to encoder it, then submits it to location code database which is realized by the HBase, through a scan operation in the database to find geological location points which are close to its location. The position
code database can quickly filter out irrelevant information, reduce the candidate set, and improve the accuracy of the information acquisition.

Content pushing service: First, the content manager receives a candidate set of geological positions provided by the LBS, at this point, each geological location contains a number of geological information; content manager's role is based on the user's interest points, using collaborative filtering and other recommendation techniques to select the information that users may be concerned about in the geological location points, so as to improve the efficiency of content pushing service.

Map Service: Data visualization background component, be responsible for presenting the pushing results to the user in a map, at the same time, completing the necessary human-computer interactive function.

The basic components of the system are composed of the content repository, the content index database, the position code database and the spatial database. The former three are built on the Hadoop ecosystem, through a number of MapReduce workflow to complete the data location information acquisition, the association mapping of geographic location-geological location-geological content, the reconstruction of the position information coding and other important data processing process. The architecture of the whole system takes the content repository [2] as the core, realized the unified description model of geological data, and it can be regarded as the data warehouse of data analysis. Content index database is the lightweight mapping of the content repository, which is easy to find and locate the data quickly.

3. Research on Key Technology
In view of the particularity of industry demand and data structure, In this paper, based on the existing location-based content pushing technology [4,5], it is necessary to seek to solve the problem of unstructured geological data automatic acquisition, organization and management, analysis and service links. In the automatic acquisition and organization of data management, this paper adopts the content tree model of document [2]. The key technologies of the analysis processing and the service links are the key focus of this study section.

3.1. Data Processing Main Flow Based on Hadoop
In this paper, data objects are the geological survey reports of rich text (text and pictures). According to the method of literature [2] to model the content of this kind of data, in this way, a geological report is expressed as a document structured based on the paragraph content item subtree, all data form a content tree with the same root node. To simplify the problem, assume that each report is initially independent of each other, that is, any two subtrees can only be connected through a path that contains the root node. Subsequent data calculations are performed on this data structure and assumptions. Figure 3 shows the data processing of this article is divided into two sub-processes, each process is composed of several MapReduce programs. Process one, position entity recognition and adjacent position coding for text data containing position information, to form primary location index, as detailed in the 2.2 section. Process two, classification with item content as the object, to form the the secondary classification index, as detailed in the 2.3 section. The primary location index and the secondary classification index are stored in the position code database.

![Figure 3. Algorithm Main Flow and Working Principle Diagram.](image-url)
3.2. Entity Recognition and Location Coding Technology

The geographical position obtained by the satellite positioning is the coordinate, and location information in geological data is mostly used in natural language text description. In order to be able to quickly and correctly push the information in the vicinity of the location of the field workers, a method is needed to establish an association from geographic location to geological location and then to geological content. Identifying named entities (geological locations) from text data, this paper uses cascaded hidden Markov algorithm [5] and builds location index by using geohash algorithm. The whole algorithm process is based on the Mapreduce technology framework; the final results are saved in the HBase, then using the geohash code as rowkey, and value column stores geological location-related content referenced in the content repository. Pseudo code:

```java
Class Mapper
   procedure Map(null, content(i))
      use stacked HMM to identify the location entity {POIentire} in content(i).text
      if count(POIentie)>0 then
         for each poi in {POIentire} do
            emit(poi, content(i).ref)
      after shuffle and sort, the contents of the same location entity are sent to a reduce program.

Class Reducer
   procedure Initialize
      // Load position < Body -Coordinate> Map
   procedure Reduce(poi-entire, Lst(content-ref))
      coords = Map(poi-entire)
      using coords to generate geohash code of poi-entire
      //take key(geo) as linekey, take the values of poi-entire and Lst(content-ref)
      //as column value write in HBase
      for val in [poi-entire, Lst(content-ref)] do
         emit(key-geo, val)
```

3.3. Geological content classification based on machine learning

The work of Section 3.1 is to associate a large number of geological content with geological location. So the reasonable classification of geological content is the key technology to realize the correct information pushing, this is the premise of the system to filter out a lot of irrelevant information. As mentioned above, the geological content is basically recorded in the form of text in the geological survey reports, and a report will contain a large number of different kinds of information, which is often divided in chapters. Due to the huge amount of report, the same kind of geological content point is very dispersive with respect to the overall information in general.

Based on the above analysis, text classification technology is a feasible way to solve this problem. But, if take the document as the smallest level of granularity, the information entropy will be larger, which is very inconvenient for establishing characteristic words and specifying the classified model. This is one of the reasons why we use a paragraph of the document as the smallest level of granularity. According to the modeling method of literature [2], content-item is persistent in HBase, can be directly used as the input of the processing algorithm.

At the same time it should be noted that, if we split each document by paragraph again, the amount of data that needs to be processed will reach one hundred million, so it needs to be done with the MapReduce computing framework. We use machine learning to accomplish the task of mass text classification. Algorithm flow can be divided into the following steps:
1: Stopping word, Chinese word segmentation, count up each sample document frequency, calculate the cross entropy of each feature word \( t \), and eventually save the largest collection of \( K \) feature words \( \{ t_1, t_2, ..., t_k \} \).

2: Training corpus, using TF-IDF [6] to build VSM model for each classification, this is a \( d \times K \) order matrix, \( d \) is the number of samples in each classification, \( K \) is the number of feature value selected in step 1, \( w(d,k) \) is the value of tf-idf.

3: Use MapReduce2 combined with TF-IDF [7,8] to set up the feature vector of the text that to be classified.

4: Use KNN algorithm which is based on MapReduce to complete the classification of text.

The main workload of processing is the establishment of the document VSM model and the iteration of the KNN algorithm, and VSM established mainly relies on cross-entropy and TF-IDF these two methods. This step appear in the training and determine period at the same time, therefore their parallelization will greatly improve the performance.

4. Application Scenarios

The system is deployed in the geological survey cloud platform, by cloud service way to provide functionality for users, the client side use mobile-preferred web front-end development framework, through the browser can use the full functionality of the system. As an application example, this paper uses the real geological survey data as the basic data source.

Somewhere, geological survey workers often focus on the relationship between the various factors and the mineralization in a specific ore-forming geological background. Assumes that field workers engaged in regional geological surveys are located somewhere in Xigaze region of Tibet, and they want to get their location and the surrounding geological information by portable handheld devices. Figure 4 shows the working interface using the system function on a panel computer; system obtains geographical coordinates \( (90.7031^\circ \text{ E}, 30.7040^\circ \text{ N}) \) of the field workers through the LBS, and through the geological content pushing service to push 5 geological locations of a total of 45 geological contents for the users. System interface is displayed in accordance with the weight of the ranking of the second geological location of the "一、构造地层区划分 等" the 5 geological contents.

![Figure 4. The User Interface of System Function.](image)

5. Conclusion

Intelligent geological survey system is one of the key tasks of "Geological cloud" construction program China Geological Survey Bureau. In this paper, combined with the actual demand of field geological work, designed geological content pushing system based on location, it is an exploration using big data technology to solve practical problems, its service model is the embodiment of the "recycling field geological survey geographic information intelligent sensing process". The next step, we will proceed with the construction of geo-information in-depth analysis and knowledge discovery using platform, to build the value chain of the geological results data, and launch geosciences’
information resource sharing and value-added services, to create favorable conditions for the information integration, resource sharing and knowledge innovation of geological data.

Acknowledgement
Financial support was provided by China geological survey (No. 1212011302660).

References
[1]  Li C L, Li J Q, Zhang H C, Gong A H and Wei D Q 2015 Big Data Application Architecture and Key Technologies of Intelligent Geological Survey Geological Bulletin of China 34 pp 1288-1299
[2]  Wei D Q, Li C L, Wumuti N, Wei J X and Yang J L 2014 Organizing and Storing Method for Large-scale Unstructured Data Set with Complex Content 2014 Fifth International Conference on Computing for Geospatial Research and Application IEEE 2014 pp 70-76
[3]  Wang X, Li Q H and Zhang F J Study on LBS-based Smart Push Technology 2011 Communications Technology 44 pp 95-97
[4]  Xiao F, Hou Y, Wang L Z and Jia B 2015 Research of Intelligence Information Push Technology Based on LBS Geomatics&Spatial Information Technology 38 pp 125-127
[5]  Yu H K, Zhang H P, Liu Q, Lv X Q and Shi S C 2006 Chinese named entity identification using cascaded hidden Markov model Journal on Communications 27 pp 87-94
[6]  Salton G, Buckley C.Term-weighting Approaches in Automatic Text Retrieval 1988 Information Processing & Management 24 pp 513-523
[7]  Chu C T, Kim S K, Lin Y A, Yu Y Y and Bradski G 2006 Mapreduce for machine learning On multicore Proceedings of Neural Information Processing Systems Conference19 Vancouver:MIT Press pp 281-288
[8]  Fang P L and Yang B R.Study on KNN arithmetic based On cluster 2009 Computer Engineering and Design 30 pp 4260-4262