Research on the Focused Crawler of Mineral Intelligence Service Based on Semantic Similarity

Shiqi Deng¹, ²
¹Computer Academy, Xiangtan University, Xiangtan, People’s Republic of China, 411105
²Hunan Key Laboratory of Internet of Things and Information Security, Xiangtan University, Xiangtan, People’s Republic of China, 411105
Email: 807684371@qq.com

Abstract. Large-scale general search engines have been unable to meet the needs of “specialized, sophisticated and deep” information in the field of mineral intelligence services. Vertical search engines have emerged at the historic moment, and the focused crawler is the key technology to achieve vertical search engines. This paper proposes a hybrid topic strategy based on text content and web link structure for the characteristics of mineral information field. In order to improve the topic relevance determination ability of the focused crawler, this article introduces HowNet and the word embedding technology in the field of natural language processing, and combines them to carry out the text-based topic relevance determination; at the same time, it also introduces the HITS algorithm based on the link structure of web pages. The topic strategy based on text content is organically combined with the topic strategy based on the webpage link structure to realize the recognition and prediction of the webpage topic. Simulation experiment results show that the method proposed in this paper can achieve a high recall rate and precision rate for the acquisition of mineral intelligence information on the Internet.

1. Introduction

There are many sources of mineral intelligence information, such as government work reports, government tender announcements, policy documents, market investment trends, news, periodicals, books and so on need to be considered; at the same time, mineral intelligence information involves a variety of forms, its content is also very complex, and most of them have a certain timeliness, need to be updated in time. At present, China's mineral intelligence information is mainly searched and sorted by manual or semi-automatic methods. Its workload is extremely huge and the cost is relatively high.

It is difficult for users to obtain accurate and professional mining information in the mass of Internet information by the traditional search engine; At the same time, as the core technology of traditional search engine, the performance and cost requirements of universal web crawler are not satisfactory in the current Internet environment. As an extension and development of the traditional search engine, the vertical search engine has the characteristics of "specialization, precision, and depth"; and the focused crawler is the core technology of the vertical search engine. In order to avoid crawling irrelevant web pages, focused crawler sort the web pages according to their topic relevance and crawl them in turn. The performance of the focused crawler is mainly determined by the topic strategy [1], which has become a current research hotspot. The main research work of this paper is to propose a topic strategy for the acquisition of mineral intelligence information on the Internet.

In view of the problem of information acquisition of geological and mineral resources, literature [2] proposed a solution that uses the TF-IDF algorithm to obtain webpage keywords and calculate topic
relevance. The application of this scheme can obtain a good precision, but since the scheme does not consider the semantic relationship of the text, the crawling text subject words must match the query key words perfectly, so that the crawling results are very limited. In terms of topic strategy based on webpage text content, a large number of scholars have published application studies in conjunction with semantic analysis technology. For example, in 2010, Z L Jiang proposed a topic strategy that combines SVM, semantic analysis technology and Bayesian [3]; In 2018, H G Sun adopted the LDA model and introduced the similarity calculation model of semantic information to design the semantic focused crawler [4].

The existing mainstream topic strategies are mainly based on three methods: based on webpage text content, based on webpage link structure, and based on mixed text content and link structure. A single judgment topic based on webpage text content lacks of globality, and there is a problem of "myopia"; a single judgment topic based on the webpage link structure is prone to "topic offset". Therefore, topic strategy based on web page link structures are mostly used in combination with topic strategy based on web page text content. For example: Q Chen and L B Luo designed a focused crawler combining HITS and Shark-Search [5]; G Y Zhang and L Li proposed to combine the HITS algorithm and SVM classifier [6]; N Xu and Q S Zhu improved the OPIC method and proposed the OPIE method [7].

Based on the aforementioned method, this paper proposes a hybrid topic strategy for intelligence gathering in the mining field. The paper first calculates the page information semantic relevancy by HowNet[8] and word embedding technology, then calculates the final access priority of link with HITS algorithm, and obtains the download queue. The mining information collection method proposed in this paper includes three modules including crawling module, storage module and post-processing module. The focused crawler based on a hybrid topic strategy of semantic analysis and link structure analysis is implemented by the crawling module. This module needs to collect and preliminarily filter the content of the mining intelligence field on a given website. The storage module will persist the content submitted by the crawl module and then pass it to the post-processing module for cleaning to obtain the final mining information. Simulation experiment results show that the method proposed in this paper is superior to the traditional existing methods in terms of recall rate and precision.

2. Method

A starting point for intelligence gathering is to put the seed URL into the crawl queue thus opening the work of the crawl module. The work of the crawl module has the following steps:

Step 1: First of all, the web pages in the crawling queue are extracted text, simplified conversion, removal of stop words and low frequency words, word segmentation and other page parsing work.

Step 2: For the parsed text, the module calculates the semantic similarity to the topic based on HowNet and the trained word vectors; and the TF-IDF model is used to calculate the word weight; finally, the page information semantic relevancy is calculated by combining semantic similarity and word weight.

Step 3: The node of the web page is added to the root set of the HITS algorithm module. When the link in the root set is sufficient, the module starts iterative calculation to obtain the Authority and Hub of the unvisited link.

Step 4: Combine the page information semantic relevancy of the parent page, the page information semantic relevancy of the anchor text of the unvisited link on the parent page, and the authority of the unvisited link, calculate the priority of the unvisited web page, and make a threshold judgment. If it is greater than the threshold, it will join the crawl queue.

The overall process design is shown in Figure 1:
3. Key Technology

3.1. Calculate Word Similarity
Word similarity is a numerical value used to describe the degree of similarity between words, and its value range is between [0, 1]. The closer to 1, the more similar the words. Conversely, the closer to 0, the more different the words.

HowNet is an important knowledge base in the field of natural language processing established by Z D Dong and Q D Dong, which explains the concepts of Chinese and English words used in life through the sememe, and links words to each other through this interpretation method, thus establishing a complex network of knowledge structure.

HowNet is compiled by manual, the accuracy is high, but the vocabulary in professional field is scarce and the update speed is slow. By using the training corpus of word embedding technique, word vector can complement HowNet, so this method proposes to combine the two.

1) Training word embedding vector; and set the subject word set T.
2) A pre-processing of crawling content includes extracting text and anchor text, word segmentation, removing stop words, etc. After processing, the word set to be matched M is obtained
3) The words in the subject word set T and the word set to be matched M are extracted from the HowNet. According to the method proposed in [9], the semantic similarity SimH between the word to be matched covered by HowNet and the subject word covered by HowNet is calculated.
4) The trained word vectors are used to represent the words in the subject word set T and the word
set to be matched \( M \), and the cosine similarity is calculated according to Formula (1).

\[
\text{SimV}(\textbf{m}, \textbf{t}) = \frac{\textbf{m} \cdot \textbf{t}}{\|\textbf{m}\| \times \|\textbf{t}\|} = \frac{\sum_{i=1}^{l} m_i t_i}{\sqrt{\sum_{i=1}^{l} m_i^2 + \sum_{i=1}^{l} t_i^2 - \sum_{i=1}^{l} m_i t_i}}
\]  

(1)

Among them, \( \text{SimV} \) represents the similarity between the word to be matched and the subject word, \( \textbf{m} \) represents the word vector to be matched, \( \textbf{t} \) represents the subject word vector, \( m_i \) and \( t_i \) respectively represent the size of word vector \( \textbf{m} \) and word vector \( \textbf{t} \) in the dimension \( i \), \( l \) represents the vector length of word vector.

5) Design parameter \( \alpha \), let \( 0 < \alpha < 1 \), calculate the final semantic similarity \( \text{Sim} \) according to Formula (2).

\[
\begin{align*}
\text{Sim} &= \text{SimV} & (\text{SimH} = 0) \\
\text{Sim} &= \alpha \text{SimH} + (1 - \alpha) \text{SimV} & (\text{SimH} \neq 0)
\end{align*}
\]  

(2)

3.2. Calculate Word Weight

Through the calculation in 3.1, the similarity between the words contained in the web page text and the selected topic word can be obtained. When calculating the page information semantic relevancy based on the text content, it is necessary to define the weight of the words participating in the calculation. Formula (3): 

\[
w_i = tf_i \cdot idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \log \frac{D}{D_t}
\]  

(3)

Among them, \( tf_i \) represents the word frequency of the word \( t_i \), \( idf_i \) represents the inverse document frequency of the word \( t_i \), \( n_{i,j} \) is the number of occurrences of the word \( t_i \) in the file \( d_j \), \( \sum_k n_{k,j} \) is the total number of occurrences of all words in the file \( d_j \), \( D \) is the total number of all files in the corpus, and \( D_t \) is the number of all files containing the word \( t_i \).

3.3. Calculate Page Information Semantic Relevancy

After finishing the calculations in 3.1 and 3.2, according to the calculation results of Formula (2) and Formula (3), we can calculate the page information semantic relevancy, the calculation method is as Formula (4):

\[
\text{Rel}(\textbf{T}, \textbf{M}) = \frac{\sum_{m=1}^{M_t} \sum_{t=1}^{T_t} w_m w_t (\text{Sim}(\text{word}_m, \text{word}_t))^2}{M_t}
\]  

(4)

Among them, \( \text{Rel}(\textbf{T}, \textbf{M}) \) represents the page information semantic relevancy between the subject word set \( \textbf{T} \) and the word set to be matched \( \textbf{M} \), \( w_m \) represents the weight of the word to be matched \( \text{word}_m \), \( w_t \) represents the weight of the subject word \( \text{word}_t \), \( M_t \) represents the size of the word set to be matched, \( T_t \) represents the size of the subject word set, \( \text{Sim}(\text{word}_m, \text{word}_t) \) represents the word similarity between \( \text{word}_m \) and \( \text{word}_t \).

3.4. Link Structure Analysis

The HITS algorithm was proposed by Kleinberg. The algorithm uses Authority and Hub (centrality) to measure web pages. The basic idea of the HITS algorithm is that the Hub and Authority of a webpage can be balanced with each other. The higher the authority of the page set pointed by a webpage, that is, the higher its own Hub value; a page is pointed to by many pages with high Hub values, then the higher its own Authority.

The HITS algorithm first determines a directed network subgraph \( G(V, E) \) (\( V \) is a node set and \( E \) is an edge set) according to the preset subject words, and then iteratively calculates the Hub and Authority of each web page. Specific steps are as follows:

1) First, the root set most relevant to the topic is obtained by manual selection.
2) The base extension set is obtained by performing link analysis on the URLs in the root set. The extension principle is that for each web page \( P \) in the root set, the link pointing to \( P \) is added to the
3. The Authority value and Hub value of each web page are set to the parameters \( a(v) \) and \( h(v) \), and are initialized with 1.

4) To calculate the Hub and Authority of each page in the base set, the algorithm is as follows:

The sum of the Hub weights of all pages linked to web page \( v \) is its Authority weight:

\[
a_i(v) = \sum h_{i-1}(w)
\]  

(5)

The Hub score of webpage \( v \) is the sum of the authority weights of the pages pointed to by webpage \( v \):

\[
h_i(v) = \sum a_{i-1}(w)
\]  

(6)

To calculate the Priority of Unvisited Pages

Finally, we apply a topic determination strategy based on semantic similarity combined with the HITS algorithm, taking into account the topic relevance of the content of the parent page and the topic relevance of the anchor text of the child page, to calculate the priority of an unvisited URL (referred to as priority). The priority will determine whether the URL needs to be crawled, and the calculation Formula is as Formula (9).

\[
p(\text{url}) = \beta \text{Rel}(T,M_{\text{parent\_page}}) + \gamma \text{Rel}(T,M_{\text{Anchor\_text}}) + (1 - \beta - \gamma) a(\text{url})
\]  

(9)

Among them, \( p(\text{url}) \) indicates the priority of the webpage, \( \text{Rel}(T,M_{\text{parent\_page}}) \) is the page information semantic relevancy of the parent page body of the unvisited page, \( \text{Rel}(T,M_{\text{Anchor\_text}}) \) is the page information semantic relevancy of the unvisited page's anchor text on the parent page and its context, \( a(\text{url}) \) represents the Authority of the URL of the unvisited page calculated by the HITS algorithm. \( \beta, \gamma \) are balance parameters, \( 0 < \beta, \gamma < 1 \). The value of the balance parameter is determined by AHP.

Whether to add unvisited URLs to the crawling queue depends on whether the priority of the webpage is greater than the threshold, and the threshold is dynamically determined by the negative feedback mechanism. If more than 130 pages are continuously below the threshold, the threshold falls, and if more than 50 pages are continuously above the threshold, the threshold rises.

4. Experimental Results and Analysis

The experiment in this paper is implemented in Python environment, using the jieba word segmentation tool to perform word segmentation on the training corpus and the crawled text; using the word2vec model in the gensim toolkit to train the word vector; using the TF-IDF model in the gensim toolkit to obtain the text weight of words.

This paper selects the data in wikipedia's chinese database and the special data of mining website as the training corpus. Wikipedia has a large amount of data, wide coverage and various categories, but it does not cover the mining field completely, and the mining special data focuses on the mining field, which complement each other. Subject word sets are selected according to mining topic data and Fudan university chinese text classification data sets to ensure professionalism. The weight of subject
words is determined based on the Chinese text classification data set of Fudan University and calculated according to Formula (3). For the words that do not appear in the Chinese text classification data set of Fudan University, as well as the non-mining field-specific words with higher weight (such as "China", "department" and other words), the weights need to be manually adjusted.

For the value of the parameter $\alpha$ in Formula (2), select 350 groups of words to perform the semantic similarity judgment experiment according to the method proposed in 3.2 (the case of $\text{Sim}_H = 0$ is not discussed), and the accuracy rate was taken as the standard. The experimental results are shown in Figure 2. Compared with using only one of the HowNet and word embedding techniques for semantic similarity analysis, the results are more accurate.

![Figure 2. The value of the parameter $\alpha$ experiment result](image)

This article selects multiple websites as shown in Table 1 for experiments. The websites are divided into three categories. The first two websites are portal websites, which contain a large amount of information in irrelevant fields such as military, finance, technology, and fashion. The third and fourth websites are related to natural resources. Among which there is more information in the field of mining, but at the same time they contain a lot of interference information such as survey, geology, etc. The last two websites belong to the field of mining, and most of their contents belong to the field of mining information.

| Website            | Name                                      |
|--------------------|-------------------------------------------|
| www.163.com        | Netease                                   |
| www.sohu.com       | Sohu                                      |
| www.mnr.gov.cn     | Ministry of Natural Resources Portal      |
| zrzyt.hunan.gov.cn | Hunan Provincial Department of Natural Resources |
| www.chinamining.org.cn | China Mining Network                      |
| www.zgkyb.com      | China Mining News                         |

In this paper, the precision and recall rate shown in Formula (10) and Formula (11) are used as the evaluation criteria:

\[
\text{Precision} = \frac{\text{relevantResult}}{\text{AllResult}} \tag{10}
\]

\[
\text{Recall Rate} = \frac{\text{relevantResult}}{\text{AllRelevantOnInternet}} \tag{11}
\]
Next, we compare the crawler based on the hybrid topic strategy with the crawler based on Shark-Search algorithm and the crawler based on HITS algorithm. The performance of the three crawler in the experiment is shown in Figure 3 and Figure 4.

As can be seen from Figure 3, the precision of the hybrid topic strategy proposed in this paper and the Shark-Search algorithm both perform well, while the accuracy of the HITS algorithm decreases significantly over time. This is because the hybrid topic strategy proposed in this paper and the Shark-Search algorithm are based on the text content to determine the topic relevance, while the HITS algorithm is based on the link structure to determine the topic relevance, the phenomenon of "topic offset" appeared in the late crawling.

It can be seen from Figure 4 that in terms of recall rate, the hybrid topic strategy proposed in this paper is significantly better than the Shark-Search algorithm, and the Shark-Search algorithm performs better than the HITS algorithm. This is because although the HITS algorithm has a small calculation amount and a fast crawling speed, there is a "topic offset" in the later period, which results in crawling a large number of irrelevant web pages and reduces the recall rate; although the Shark-Search algorithm has more accurate topic determination, it is easy to fall into the misunderstanding of "myopia".

From the above experimental results, it can be seen that the hybrid topic strategy proposed by the HITS algorithm combined with semantic analysis technology can overcome the "topic offset" problem of the topic strategy based on the link structure, and can also avoid falling into the "myopia" trap of the topic strategy based on text content. The method proposed in this paper can determine the potential value of links more accurately by link structure analysis on the basis of better accuracy of topic correlation judgment.

After crawling the mining information of the experiment website according to the method of this article, the precision of each website is obtained as shown in Table 2. On the website of the pure mining industry, the average precision is 98.97%; on the portal website of the natural resources field, the average precision is 81.02%; on the portal, the average precision is 69.54%.

From the experimental results in Table 2, it can be observed that the precision rate decreases in turn from pure mining domain websites to natural resource portals, and then to large portals. The reason for this result is that, on the one hand, according to the order of the above, the topic has become more and more irrelevant, and the noise is getting louder; on the other hand, its page structure is getting richer and richer, and the link relationship is more complicated.
Table 2. Experimental results

| Website                        | Corpus (Pieces) | Related number (Pieces) | Precision (%) |
|-------------------------------|-----------------|-------------------------|---------------|
| Netease                       | 3842            | 2591                    | 67.43         |
| Sohu                          | 2756            | 1974                    | 71.64         |
| Ministry of Natural Resources Portal | 3606            | 2812                    | 77.98         |
| Hunan Provincial Department of Natural Resources | 458            | 385                     | 84.06         |
| China Mining News             | 7174            | 7113                    | 99.16         |
| China Mining Network          | 6857            | 6773                    | 98.78         |

5. Conclusions and Prospects
In the Internet age of information explosion, improving search technology and improving search performance is of great significance for the integration and sharing of mineral information. Based on the characteristics of the mineral intelligence service field, this paper proposes a hybrid topic strategy combining semantic analysis technology and HITS algorithm. Based on the existing research results, this method solves the problems of "topic offset" and "local optimal solution", thereby improving the service search ability. This method has been applied to the retrieval of mineral information services and has achieved good experimental results. The next step will be to optimize the performance of the method on the portal site with low accuracy.

6. Acknowledgement
This research was partly supported by the National Natural Science Foundation of China (Nos. 61873221 and 61672447) and the Natural Science Foundation of Hunan Province (Nos. 2018JJ4058, 2019JJ70010, and 2017JJ5036). Publication costs were funded by the National Natural Science Foundation of China (Nos. 61873221 and 61672447).

The authors sincerely thank all the teachers and students who participated in this study for their guidance and help.

7. References
[1] L Qiu, Y S Lou, M Chang, "An Improved Shark-Search Algorithm for Theme Crawler", J. Microcomputer applications, vol. 2, pp. 19-21, 2017.
[2] X Duan. "Information Acquisition Technology of Geology and Mineral Resources Based on Web", J. Technological innovation and productivity, vol. 3, pp. 38-41, 2017.
[3] Z L Jiang, X Y Tian, X Zhao. "A Topic Crawler Algorithm Based on Semantic Analysis", J. Computer Engineering and Science, vol. 9, pp. 145-147, 2010.
[4] H G Sun, R Q Zang, C D Ji, etc. "Study on the algorithm of focused crawler based on semantic similarity", J. Journal of Northeast Normal University: Natural Science Edition, vol. 2, pp. 56-62, 2018.
[5] L B Luo, Q Chen, Q X Wu. "Research on Topical Crawler of Shark-Search Algorithm and Hits Algorithm", J. Computer Technology and Development, vol. 11, pp. 76-79, 2010.
[6] L Li, G Y Zhang, W W Li. "Research on Focused Crawling Technology Based on SVM", J. computer science, vol. 2, pp. 118-122, 2015.
[7] Q S Zhu, N Xu, Y Zhou. "An Adaptive Focused Crawling Algorithm Based on Link and Content Analysis", J. Computer and Modernization, vol. 9, pp. 77-80+89, 2015.
[8] Z D Dong. HowNet. http://www.keenage.com, 2013.
[9] Q Liu, S J Li. "Calculation of Lexical Semantic Similarity Based on HowNet". Chinese Computational Linguistics, vol. 2, pp. 59-76, 2002.