Design of topic Web crawler based on improved PageRank algorithm

Linxuan Yu¹, Yeli Li², Qingtao Zeng³*
¹ Beijing Institute of Graphic Communication, BeiJing, China
² Beijing Institute of Graphic Communication, BeiJing, China
³ Beijing Institute of Graphic Communication, BeiJing, China
*Corresponding author’s e-mail: zengqingtao@bigc.edu.cn

Abstract. With the continuous development of network information technology, the network is filled with a large number of all kinds of unstructured data called big data. However, this data is not easily stored in a local database. People realize that it is essential to get useful information from the Internet efficiently. The effort to gather information by human hands has led to the emergence of web crawler technology. However, the existing search engines still have shortcomings in topic similarity judgment and web page sorting algorithm. Therefore, this paper applies PageRank algorithm to topic crawler, constructs a vertical search engine, and introduces topic relevance factor to suppress “topic drift” according to the shortcomings of PageRank algorithm.

1. Introduction

With the continuous development of the society, we are using intelligent devices all the time, and taking advantage of the fragmented time to constantly receive various information from the Internet. The convenient and fast way of life affects people's work, life, entertainment and other aspects to varying degrees. According to the report published by China Internet Network Information Center (CNNIC) in 2016, [1]: by December 2015, the number of Chinese netizens has reached 688 million, and the Internet penetration rate has exceeded 50%. These data clearly show that more and more people use the Internet to serve themselves. In general, people often use search engines when using the Internet, which can help people quickly retrieve the information they need.

2. Research on crawler related technologies

2.1 Working principle of the crawler

Vertical search engine is mainly composed of the following three modules: topic crawler module, pre-processing module and query pool module. The topic crawler module provides the information source and target topic for the system, while the other modules constitute the full-text searcher. The pre-processing module and query pool provide the retrieval service and finally face to users.

The working principle is as follows: the theme web crawler crawls the URL related to the theme in the Internet through the preset theme. After the URL is downloaded, it enters the pre-processing module to extract the downloaded URL content, and the extracted information is indexed by the index program. After processing the preset keywords, the search program in the crawler module will enter into the index database for matching, and feed back the contents that are highly relevant to users' needs [2].
2.2 Architecture of topic Web crawler

Vertical search engine is mainly composed of the following three modules: topic crawler module, preprocessing module and query pool module. The topic crawler module provides the information source and target topic for the system, while the other modules constitute the full-text searcher. The preprocessing module and query pool provide the retrieval service and finally face to users.

1>download module. For the crawler, the main task is to download the web page. There is also a link scheduling module in the download module that retrieves the link from the queue to be fetched and starts the download based on the search policy.

2> theme-related content extraction. After the topic web crawler is connected, the first step is to parse the web page and match the relevant links with the pattern matching method. The analysis of web page content according to the theme has a great impact on the subsequent analysis. Denoising is an important link in the extraction of theme content. Since the noise content has nothing to do with the theme, the relationship between the extracted content and the theme is not clear enough, which affects the effect of theme identification. Besides denoising, it also includes Chinese participle, stop-word deletion and so on.

3>identifies theme pages. To identify the theme page is to determine whether the crawled content of the page is related to the theme and whether the relevance can reach the threshold. From the perspective of Chinese semantics, topics can be concepts, words, phrases, paragraphs or articles. The selection of a preset theme is the first step to extract the theme information. The theme of a web page is represented by features, which represent the theme of a web page [3].

4> extracts the relevance of linked topics and evaluates them. First remove the AD links, convert the relative links to absolute links, and then evaluate the relevance of the links to the topic and put them into the query pool. The correlation of links is primarily concerned with the correlation of the parent page and the link anchor text.

3. PageRank algorithm application and improvement

3.1. Introduction to the Algorithm

The PageRank algorithm was proposed by Larry Page and Sergey Brin, the founders of Google, in 1988. PageRank algorithm will first determine the PageRank value of each page, and then according to the size of the value of the page sort. If the link in page X points to page Y, as page X voted for page Y, if there are many page links pointing to Y, it means that page Y is very important, page Y PageRank value is relatively high [5]. PageRank algorithm not only considers the number of links on a page, but also analyzes the importance of the page itself. A more important page link, then its own importance will be higher.

PageRank algorithm is based on the following two premises:

(1) If a web page is linked by other web pages for many times, it means that this web page is very important; If a page is not linked many times, but is linked to a very important page, then the page may also be very important; The importance of a web page is equally divided among the other pages.

(2) Assume that the user browses a page in any set of browsing web pages and then links outward according to the page.

Then continue to browse, and did not return to browse the page, then browse the next page of the probability is the PageRank value of the page being browsed.

The PageRank value of page links is affected by the following three factors: the number of pages linked into the page, the page PR value linked into the page and the number of pages linked out. Therefore, the calculation method of PageRank value of webpage is as follows:

Firstly, calculate the PR value of all links of X web pages into Y web pages, divide this value by the number of links out of Y web pages L (Y), Then add up the calculated results, as shown in Equation (1).

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$ (1)
Among them, \( PR(A) \) represents PageRank value of webpage A, \( PR(A) \) represents PageRank value of webpage A, \( L(B), L(C), L(D) \) respectively represents the number of out links of webpage B, C and D. Because there will be some isolated web pages do not link to any other web pages, so add a damping coefficient \( d \) to modify formula (1), \( d \) value of 0.85, indicating that users will at any time with a probability \( d \) from the current web page hyperlink evenly browse its out of the link web page. Therefore, the exact expression (2) after the PageRank algorithm is modified is as follows:

\[
PR(A) = (1 - d) + d \left( \frac{PR(T_1)}{L(T_1)} \right) + \ldots + \left( \frac{PR(T_n)}{L(T_n)} \right) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(T_i)}{L(T_i)}
\]

The working principle of PageRank algorithm is shown in Figure 1. The web page in the Internet is regarded as a directed graph, in which the web page is the node, as shown in Figure 1, a website is composed of four web pages A, B, C, D, and they link to each other, A links to B, C links to A, D, C links to D, and D links to C.

![Figure 1. A network directed graph with four pages](image)

The square matrix form of \( 4 \times 4 \) can be used to represent the link relationship in the digraph. If there is a link from node I to node J, then \( H_{ij}=1 \), otherwise it is 0. Then the \( H \) matrix is shown in formula (3):

\[
H = \begin{bmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{bmatrix}
\]

The probability transfer matrix \( M \) of PageRank algorithm, represents the probability of node I chain to node J, the sum of each column of the \( M \) matrix is 1. Then, the \( M \) matrix is shown in Equation (4):

\[
M = \begin{bmatrix}
0 & \frac{1}{2} & 1 & 1 \\
\frac{1}{3} & 0 & 0 & 1 \\
\frac{1}{3} & 0 & 0 & 1 \\
\frac{1}{3} & \frac{1}{2} & 1 & 0
\end{bmatrix}
\]

Let the initial PR value of four pages A, B, C and D be 1, we can get the following equation:

\[
\begin{align*}
PR(A) &= 0.15 + 0.85 \times \frac{PR(B)}{2} \\
PR(B) &= 0.15 + 0.85 \times \frac{PR(A)}{3} \\
PR(C) &= 0.15 + 0.85 \times \frac{PR(A)/3 + PR(D)}{3} \\
PR(D) &= 0.15 + 0.85 \times \frac{PR(A)/3 + PR(B)/2 + PR(C)}{3}
\end{align*}
\]

After one iteration, \( PR(A)=0.575, PR(B)=0.312916, PR(C)=1.162916, PR(D)=1.434385 \). According to the calculation of the above iteration, the iteration times and results are shown in Table1:
Table 1. Iterate

| The number of iterations | PR (A) | PR (B) | PR (C) | PR (D) |
|--------------------------|--------|--------|--------|--------|
| 0                        | 1      | 1      | 1      | 1      |
| 1                        | 0.575  | 0.342916667 | 1.162916667 | 1.434385417 |
| 2                        | 0.282929583 | 0.230180382 | 1.449407986 | 1.560003832 |
| 3                        | 0.247826662 | 0.220217554 | 1.546220812 | 1.628097705 |
| 4                        | 0.243592416 | 0.219017864 | 1.602900913 | 1.674566232 |
| 5                        | 0.243082592 | 0.218873401 | 1.642254698 | 1.70781109 |
| 6                        | 0.243021195 | 0.218856005 | 1.670495432 | 1.731790925 |

After iterative calculation, until each page PR value close to a fixed value end. According to Table 2, it can be found that D has the highest Page Rank. Then compare the number of external links and PageRank value, as shown in Table 3:

Table 2. PageRank value comparison table between the number of external links and PageRank value

| Inbound links number | Export link number | PageRank value |
|----------------------|--------------------|---------------|
| D A/B/C 3            | C 1                | 1.794205385   |
| C A/D 2              | D 1                | 1.743928201   |
| A B 1                | B/C/D 3            | 0.218853624   |
| B A 1                | A/D 2              | 0.24301279    |

Given the spontaneity and disorder of the Internet link structure, the PageRank value of web pages in the calculation process may appear such a phenomenon: a group of web pages linked to each other, but these pages do not link out of any group of web pages. In this case, once there is an external webpage link to the page in the group, PageRank value will always stay in the inside of this group of pages and cannot be sent out, which is PageRank value precipitation phenomenon [6]. As shown in Figure 2:

Fig2. PageRank value precipitation phenomenon

In order to deal with the precipitation phenomenon of PageRank value, the decay factor E is introduced, and E corresponds to the initial value of PageRank. The improved formula is shown as (5):

\[
PR(A) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)} + dE(A)
\]

3.2. Shortcomings of PageRank algorithm

PageRank algorithm is to use the link structure of web pages to calculate the PageRank value, so as to list the search results, with a relatively high response speed, but the PageRank algorithm in the process of application still showed some problems:

(1) Lay particular stress on old web pages.
As can be seen from the PageRank calculation formula, PageRank value and page release or update time has nothing to do. In general, the old page compared to the new page on the Internet for a longer time, so it is more likely to be linked to the rest of the page, and the new page lack of upstream links, so the PageRank algorithm in the process of calculation, the PageRank value of these new pages will be relatively low, and this is very unreasonable.

(2) Authority of average page links.
In the PageRank algorithm in the calculation process of the page itself PR value is evenly distributed to the page out of the chain of all pages, and there is no link out of the authority of the page to distinguish. And the quality of web pages in the network is uneven, even if it is to point to the same page in each link, the quality level will be very different, so many high-quality pages get the weight and the value of the page itself is not consistent.

(3) Theme drift.
PageRank algorithm calculation only based on the hyperlink structure between pages, that is, there is no way to determine the relevance of web content, which is likely to cause a lot of low correlation pages because there are a lot of link relations and PageRank value is very high, this will cause the sorting results and user input query keywords irrelevant.

(4) Ignore users' browsing interests.
A page can be accessed twice by the user, largely depends on the user's browsing interest, and PageRank algorithm in the calculation but ignored this point.

4. PageRank algorithm improvement
In view of some problems existing in PageRank, the following improvements are made in this paper.

(1) Generally speaking, there are often a lot of new pages on the Internet that no one is interested in. The contents of these pages are usually the latest information or the latest official notices. To some extent, these pages are of high use value. But PageRank algorithm in the process of iterative calculation only consider the link structure, and most of the "old web page" content on the network old, but the number of its references is too much, resulting in the PageRank algorithm exists in the bias "old web page" problem. In addition, in general, a recent update of the content of a web page is often of relatively high value, PageRank algorithm has not been explored in this case, resulting in the PR value of the page will not be improved because of the content update. Therefore, based on the previous formula, this paper introduces the time feedback factor W(Ti) to improve the importance of these new web pages and speed up the delivery of these high-quality web pages. Time feedback factor W(Ti) is shown in Equation (6):

\[
W(T_i) = e^{-\lambda(T_i - t)}
\] (6)

Wherein, T represents the specific time required for the crawler to traverse the inventory page list of the entire search engine, t represents the time when the page is visited by the crawler, ti represents the latest update time of the page, and t-ti represents the update cycle of the page. The algorithm idea of introducing time feedback factor is as follows: the update time of adding the web page and the time required for the crawler to traverse the entire inventory web page list as the feedback factor. If the update cycle is longer, it means that the web page is more "obsolete" and the corresponding time feedback factor will be smaller. So that for a long time did not update the "old page" precipitation down. \( \lambda \) is the regulator of the time feedback factor, and adjusting\( \lambda \) changes the PageRank effect of the time feedback factor. After adding the time feedback factor, the calculation formula of PageRank algorithm is shown in Equation (7):

\[
PR(A) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(T_i)W(T_i)}{L(T_i)}
\] (7)

(2) In view of the phenomenon of topic drift in PageRank algorithm, this paper introduces topic relevance factor based on the correlation between target web pages and query words as well as the correlation between target web pages and their linked pages, so as to improve the weight of pages with high relevance to search topics.
Assume there is target page \( P_1 \) and link to his source page collection \( P_2=\{T_1,T_2,\cdots,T_n\} \). The query word entered by the user is \( P_3 \). After quantization of target page \( P_1 \), source page collection \( P_2 \) and query word \( P_3 \), it can be expressed as: \( P_1=(W_{11},W_{12},\cdots,W_{1n}) \), \( P_2=(W_{21},W_{22},\cdots,W_{2n}) \), \( P_3=(W_{31},W_{32},\cdots,W_{3n}) \).

In this paper, the spatial vector model is used to calculate the similarity between target page \( P_1 \) and user query word \( P_3 \), as shown in Equation (8):

\[
sim(P_1, P_3) = \frac{\sum_{j=1}^{n} W_{1j} \times W_{3j}}{\sqrt{\sum_{j=1}^{n} W_{1j}^2} \sqrt{\sum_{j=1}^{n} W_{3j}^2}}
\]  

(8)

The similarity between target page \( P_1 \) and its outlet page \( P_2 \) is shown in Equation (9):

\[
sim(P_1, P_2) = \frac{\sum_{j=1}^{n} W_{1j} \times W_{2j}}{\sqrt{\sum_{j=1}^{n} W_{1j}^2} \sqrt{\sum_{j=1}^{n} W_{2j}^2}}
\]  

(9)

Based on equation (8) and (9), the topic relevance factor \( S \) is proposed, as shown in Equation (10):

\[
S = \alpha_1 \text{sim}(P_1, P_2) + \alpha_2 \text{sim}(P_1, P_2)
\]  

(10)

\( \alpha_1 \) is the weight of similarity between the target page and the query word, \( \alpha_2 \) is the weight of similarity between the target page and the original page, and \( \alpha_1+\alpha_2=1 \). Here, set \( \alpha_1=0.6 \) and \( \alpha_2=0.4 \).

Considering the above three factors, the calculation of the improved PageRank algorithm is shown in Formula (11):

\[
PR(A) = (1-d) + d \sum_{i=1}^{n} PR(T_i) \times W(T_i) \times F(T_i) \times S
\]  

(11)

Algorithm process

Through the study of algorithm improvement, the improved algorithm flow can be obtained as follows:

1. Preprocess the crawler after it reaches the web page. The contents of the anchor text are read and saved.
2. Analyze the link structure between web pages, get the number of links in and out of web pages and the content of hyperlinks, and calculate the authoritative feedback factor.
3. The VSM model is used to calculate the similarity between the target page and user query words as well as the similarity between target page \( P \) and source page \( P_i \), so as to obtain the topic relevance factor.
4. Calculate the time parameters of the web page and get the time feedback factor.
5. The above three factors are introduced into the traditional PageRank algorithm, and the page is iteratively calculated PageRank.

5. Conclusion

PageRank algorithm is a sorting algorithm based on link structure, which USES feature vectors as theoretical basis and convergence basis [7]. Using PageRank algorithm to determine the importance of web pages is a general method to evaluate the importance of academic papers, according to the hyperlink structure of the network to assign importance to the number of levels. PageRank algorithm not only considers the number of page references, but also analyzes the importance of the page itself.

From the point of view of processing efficiency, PageRank algorithm is in the offline state in advance iteration calculated the PageRank value of each page, so, the user in the search process does not have to wait for a long time. PageRank algorithm has advantages in computing efficiency, which makes it more suitable for the current topic Web crawler algorithm, so PageRank algorithm is used in web crawler.
And for PageRank algorithm biased to old pages, average distribution weight, topic drift and other shortcomings, the introduction of time feedback factor, improve the score of "new" pages, an authoritative feedback factor is introduced to improve the weight of the linked web pages, a topic relevance factor is introduced to suppress topic drift.

Acknowledgements
(1) Major Science and Technology Projects of Guangdong Province in 2019, No. 190826175545233
(2) Beijing science and technology innovation service capability construction project (PXM2016_014223_000025)
(3) BIGC Project (Ec20207)

References
[1] China Internet Network Information Center. Statistical Report on Internet Development in China (the Fourth)[R]. 2017
[2] Rungsawang A, AngKawattanawit N. (2005) Learnable topic-specific web crawler. Journal of Network&Computer Applications, 28(2):97-114. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
[3] Wang Fang, Chen Haijian. In-depth analysis of the key principles of Web topic crawler [J]. Microcomputer Applications, 2011, 27(07): 32-34
[4] Liu Jian, TANG Huifeng, LIU Wuying, et al. A method for Extracting Chinese terms based on statistical Techniques [J]. Chinese Journal of Science and Technology, 2014, 16(05): 10-14
[5] Gao Guoshun, Chen Junhua. Research on Page-based Information Retrieval Sorting Algorithm [J]. Computer and Digital Engineering. 2011, 39(02): 47-51
[6] Li Yibing. Research on Web page sorting Algorithm based on Search engine [D]. Shenyang: Shenyang University of Technology, 2011
[7] Wan Xiaosong, WANG Zhihai, Yuan Jidong. A Heuristic algorithm for paper index ranking based on sparse matrix [J]. Journal of Computer Applications, 2015, 35(10): 2733-2736, 2741