Research on Decision Strategy Algorithms of Web Crawler Vector Space Model Based on Schmidt Orthogonal Optimization

Mu Yang [1, a] and Yanmei Hu [2,b,*]

[1] Chengdu Medical College,Chengdu,Sichuan, 610083,China

[2] Chengdu Medical College,Chengdu,Sichuan, 610083,China

[a]udjtrt@126.com; [b]13965808@qq.com

*corresponding author

Keywords: Search engines; Schmidt orthogonal; Vector space model; Text classification

Abstract. In this paper, a decision strategy algorithm for vector space model of web crawler based on Schmidt orthogonal optimization is constructed. The algorithm uses Schmidt method to orthogonally optimize the classified vectors in the vector space model, corrects the base coordinate axis of small angle projection by orthogonal optimization, realizes the rotation of vectors in the coordinate axis, and removes the document vector. The influence of relevance on classification can improve the retrieval accuracy. KNN algorithm is used to classify the orthogonal optimized document vectors. Experiments show that this method can greatly eliminate the influence of correlation on retrieval results in the process of document classification of web crawler, and improve the accuracy of document classification.

Introduction

The wide application of search engine technology provides an effective way for users to obtain effective information on the Internet. At present, the commonly used search engines are Goolge, Baidu, Sohu and so on. Search engines usually use one or more resource collectors to collect valuable data from the Internet, then index these data on local servers, and provide users with keyword-based information query services. Search engine is also called web spider, web robot, web crawler and so on.

Although the crawler has realized the automatic collection of information on the Internet, the collected data often involve too many fields and the amount of information is still too large. When users retrieve information, the results are not accurate or the response efficiency is too low, and the results are still unsatisfactory. To solve these problems, we propose a decision strategy algorithm based on Schmidt orthogonal optimization for web crawler vector space model. Network crawler decision strategy algorithm can effectively reduce the number of resources collected, improve the subject regularity of the collected resources, improve the utilization of network bandwidth, and improve the efficiency of information retrieval. It is a very useful solution for automatic collection of network information resources.

Universal Network Crawler System

The general network crawler structure mainly includes the following important functional modules: URL waiting queue, URL weight determination module, page collection module, URL extraction module, page storage and index module.The working steps of the web crawler are as follows:

1) Put the URL seed in the waiting queue of the URL.
2) Extract the URL from the URL queue to determine whether the URL has been crawled. If the queue is empty, the crawl ends.
3) If the URL has not been crawled, crawl the page it refers to, and repeat step 2 if it fails.
4) Extract the URL from the page, put it in the URL queue, and store the page in the repository.
5) Repeat steps 2 to 4 until the crawling termination conditions are met (working time, maximum number of URLs collected, fatal errors, etc.).

  DOI: 10.25236/ciais.2019.017

Decision Strategy of Web Crawler Vector Space Model Based on Schmidt Orthogonal Optimization

Schmidt Orthogonal Optimization

Schmidt method is a method to construct a set of standard orthogonal vectors from a set of linearly independent vectors. Let $\alpha_1, \alpha_2, \cdots, \alpha_r$ be a linear independent column vector in Euclidean space. Now we find a standard orthogonal basis in the dimensional linear subspace generated by this column vector. It is carried out in two steps:

I. Orthogonalization

Let $\beta_1 = \alpha_1$

$$\beta_2 = \alpha_2 - \frac{(\alpha_2, \beta_1)}{(\beta_1, \beta_1)}\beta_1$$

$$\beta_3 = \alpha_3 - \frac{(\alpha_3, \beta_1)}{(\beta_1, \beta_1)}\beta_1 - \frac{(\alpha_3, \beta_2)}{(\beta_2, \beta_2)}\beta_2$$

$$\beta_r = \alpha_r - \frac{(\alpha_r, \beta_1)}{(\beta_1, \beta_1)}\beta_1 - \frac{(\alpha_r, \beta_2)}{(\beta_2, \beta_2)}\beta_2 - \cdots - \frac{(\alpha_r, \beta_{r-1})}{(\beta_{r-1}, \beta_{r-1})}\beta_{r-1}$$

It is easy to verify $\beta_1, \beta_2, \cdots, \beta_r$ is orthogonal vector group.

II. Unitization

Let $\nu_1 = \frac{\beta_1}{\| \beta_1 \|}, \nu_2 = \frac{\beta_2}{\| \beta_2 \|}, \cdots, \nu_r = \frac{\beta_r}{\| \beta_r \|}$

Obviously, $\nu_1, \nu_2, \cdots, \nu_r$ is a group of standard orthogonal vectors, which is a standard orthogonal basis of subspace $span\{\alpha_1, \alpha_2, \cdots, \alpha_r\}$.

Starting from any set of bases $\alpha_1, \alpha_2, \cdots, \alpha_r$ in the r-dimensional product space, a standard orthogonal basis can be constructed by Schmidt orthogonalization method.

We adopt QR decomposition:

Let $A \in C_n^{m \times n}$,then A can be uniquely decomposed into

$A = UR$ or $A = R_1 U_1$

Among them, $U, U_1 \in C^{m \times n}$, R is a triangular array in the front line and R1 is a triangular array in the bottom line. (That is, the elements on the main diagonal of R and R1 are all positive).

Let $A \in C_r^{m \times r}$ (called column full rank matrix),then A can be uniquely decomposed into

$A = UR$

Among them, $U \in C_r^{m \times r}$, R is a triangular array of order r.

Vector Space Model

Vector space model is a statistical model about document representation. Vector space model is based on the key assumption that the order of entries in an article is irrelevant, and that they play an

independent role in the category of a document. Therefore, a document can be regarded as a set of disordered entries. In this model, feature items are used as coordinates of document representation, and documents are represented as a point in multi-dimensional space in the form of vectors. For example, in the k-dimensional binary vector space model, we represent a document as follows:

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{ik}) \quad d_{ij} = 0 \text{ or } 1$$

d_{ij} is the weight of the feature in the document. Figure 1 illustrates the vector space model of the document:

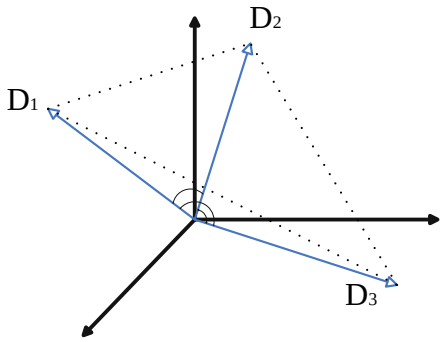

Figure 1. Document Vector Space Model

The three thick black arrows in the figure constitute a three-dimensional coordinate space, and the thin hollow arrows are represented by documents and three-dimensional vectors respectively. It can be seen from this that if the feature orientation is selected properly and the feature statistics strategy is reasonable, we can use quantified methods to classify documents on different topics.

KNN Classification Algorithms

KNN algorithm is used to calculate the cosine of the angle between the feature vectors of the samples in the complete set of documents to be classified and training samples respectively. K neighbor samples nearest to the feature vectors of documents to be classified are selected. Among them, the category with the largest proportion of candidate categories in the K neighbors is used as the category of documents to be classified, and the class like public is used as the category of documents to be classified. Formula calculates the KNN algorithm score of the page.

$$scoreKNN = \begin{cases} \dfrac{(p-n)\times m_p}{N}, & p > n \\ 0, & p = n \\ \dfrac{-(n-p)\times m_n}{N}, & n > p \end{cases}$$

Implementation and Performance Test of Decision Strategy Algorithms for Web Crawler Vector Space Model Based on Schmidt Orthogonal Optimization

After using the decision strategy of vector space model of network crawler based on Schmidt orthogonal optimization, when using algebraic features to classify information, the classified feature vectors are orthogonal because of Schmidt method optimization. Especially when the number of vectors is large, the selected feature vectors can be rotated by coordinate axis, and the correction is small. The angle projection base coordinate axis removes the correlation of the selected projection feature vectors, improves the classification accuracy, and improves the crawling accuracy of the system. System data initialization status information is shown in Figure 2. Statistical information of system operation results is shown in Figure 3. Concurrent performance testing is shown in Figure 4.

Figure 2. System Data Initialization Status Information

Figure 3. Statistical Information of System Operation Results

The system performance test results is shown in Table 1. System accuracy running test Results is shown in Table 2.

Table 1. Concurrent performance testing

Number of concurrent threads	Creeping time	Page collection number	Acquisition rate
10	1hour	2247	154.12/min
20	1 hour	5973	416.22/min
30	1 hour	932	882.88/min

Table 2. System accuracy running testing

Before or after optimizing algorithms	Total number of URL page downloads	Number of K1 pages	Number of K2 pages	Accuracy rate
Before	8742	4210	4532	48.16%
Before	6780	1288	5492	19%
After	8742	6213	2529	71.07%
After	6780	4210	2572	62.09%

Conclusions

Search engine is an important way to obtain information. Web crawler is the most important part of search engine, and its design quality directly affects the function of topic search engine. This paper

designs and implements a decision strategy algorithm system of web crawler vector space model based on Schmidt orthogonal optimization. With efficient design strategy, optimized and improved, related pages can be quickly crawled on the Web. Through a lot of experiments and tests, the results show that the algorithm has better performance and can crawl to high-quality web pages accurately.

Acknowledgements

This project was supported by Education and Teaching Research Project of Chengdu Medical College (No. JG201734) and Research Projects of Sichuan Education Information Application and Development Research Center(No. JYXX15-009)

References

[1] Menshchikov A , Komarova A , Gatchin Y , et al. A study of different web-crawler behaviour[C]// 2017 20th Conference of Open Innovations Association (FRUCT). IEEE, 2017.
[2] Elaraby M E , Moftah H M , Abuelenin S M , et al. Elastic Web Crawler Service-Oriented Architecture Over Cloud Computing[J]. Arabian Journal for Science & Engineering, 2018.
[3] Kapoor A , Arora V . Application of Bloom Filter for Duplicate URL Detection in a Web Crawler[C]// 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC). IEEE, 2016.
[4] Kumar M , Bindal A , Gautam R , et al. Keyword query based focused Web crawler[J]. Procedia Computer Science, 2018, 125:584-590.
[5] Catalin M , Cristian A . [IEEE 2017 21st International Conference on System Theory, Control and Computing (ICSTCC) - Sinaia (2017.10.19-2017.10.21)] 2017 21st International Conference on System Theory, Control and Computing (ICSTCC) - An efficient method in pre-processing phase of mining suspicious web crawlers[J]. 2017:272-277.
[6] Li Y . Research on Hotspot Topic Discovery Algorithm Based on Web Mining Technology[C]// 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). IEEE, 2017.
[7] Wei T M , Ning S X , Xun J , et al. The RESTful web services and knowledge base collaborative driven real-time tracking of emergency network opinion[J]. Journal of Shandong University, 2017.