Research on Polynomial Regression Prefetching Model

Li Chun-jie¹, Li Hong-bo²*, Li Hua¹, Zhang Li-ming¹, Fu Chun-yan¹, Guo Ji-ping¹

¹Department of Information and Electronic Technology, Jiamusi University, Jiamusi, Heilongjiang Province, 154007, China
²Department of Materials Science and Engineering, Jiamusi University, Jiamusi, Heilongjiang Province, 154007, China

*Corresponding author’s e-mail: jiamusilihongbo@163.com

Abstract. Based on the addition of prefetching function in the Web, the polynomial regression prefetching model is studied, it designed to mitigate low query response times due to network latency. In the prediction module, the polynomial regression technique is used to predict the prefetch probability of each cache item in the semantic cache. In the prefetch module, a prefetch queue is generated. When the N+1th access period comes, the system idle time is used for prefetching to reduce the web access delay. Through the test and analysis of the Web query data added to the polynomial regression prefetch model, the exactness and stability of the polynomial regression prefetch model are proved, which can reduce the network delay and improve the query response speed.

1. Introduction
The contradiction between the rapid development of the Internet and the inherent delay of the network is increasing. As the network load increases and the network congestion worsens, the quality of service that users should receive is seriously affected. At present, the use of prefetching technology to slow down network delay has become a mainstream prevention and control method. The essence of prefetching is actually delay tolerance technology. Web prefetching (prefetching) is an effective method to solve Internet network delay and improve quality of service (QoS) [1]. Web prefetching models are generally based on predictive algorithms to obtain Web objects in advance. This technology takes full advantage of the idle time of the I/O system and reduces network latency when users access it. The quality of prefetching technology is mainly measured by two performance indicators: accuracy rate and recall rate [2].

2. Prefetch model

2.1. Prefetch
The prefetching technology is divided into real-time prefetching technology and non-real-time prefetching technology. These two technologies are applied to the client and proxy servers respectively. Prefetching is performed based on the user's previous access record and the currently viewed web page, and then the predicted web page is retrieved locally from the server. Non-real-time prefetching is based on the user's historical access record. The proxy server prefetches from the web server to avoid the impact of server peak performance on server performance. Prefetching is classified according to the prefetch algorithm and can be divided into two types: access-based and semantic-
based prefetching. The prediction model established according to the user's access path is called a prefetch model based on the access path; the prefetch model for predicting the relationship between user request semantics according to the description text on the hyperlink and predicting the user request according to the relationship, called a semantic-based prefetch model[3-5]. Pre-fetching based on access paths is very effective for access patterns with strong path dependencies, but cannot predict documents that have never been accessed; semantic-based prefetching can predict web pages that users have never visited. It takes semantic-based prefetching in the Web, which is suitable for many types of prefetching services[6].

2.2. Semantic Cache
Semantic caching is the caching of query results and related semantic information, using semantics to provide data for future queries. The theoretical basis of the semantic cache application is the semantic locality of the user query. The cached content is usually the result of a (select-project-join, SPJ) query. Compared with other technologies, the semantic cache is transparent to users and does not need to change existing services; it has certain self-management capabilities, which can reduce manual management costs; more importantly, it uses query semantics to organize data with greater flexibility[7].

2.3. Polynomial regression prefetch model
The basic framework of the Web prefetch model consists of the Web log Log/Web Page, the prediction module, the Prefetching Module, the cache cache, and the standard Evaluation Criteria. As shown in Figure 1. A predictive model is built based on the information the user visits the web page. In the model, the prediction candidate is obtained by a certain prediction criterion, and the prediction module is responsible for predicting the access frequency of each semantic cache item in the next cycle; the prefetch module is based on the predicted value of the access probability of each semantic cache item, and the prefetch threshold. The prefetch queue is generated by the prefetch flag and the cache coherency valid time. The proxy server prefetches according to the prefetch queue, saves the prefetch result in the valid data storage area of the outer cache, and prefetches the semantic cache in the queue. The description item is saved in the effective semantic cache of the inner cache. Finally, the performance parameters are fed back to the prediction model for more accurate prediction.

3. The key technology of polynomial regression prefetching model
The prefetch performance evaluation index includes the accuracy (precision), which can reflect the correctness of the prediction model; the recall rate is used to measure the applicability of the prediction model, and the network traffic increase rate ITR. In the pre-fetch performance evaluation index, the relationship between the accuracy rate, the recall rate and the network traffic increase rate is that as the pre-fetch accuracy rate increases, the recall rate decreases, and the network traffic increase rate also follows reduced[8-10]. The study of prefetching modules has important implications for predictive performance evaluation. Among the components of the polynomial regression prefetch model, the most critical technology is the prediction module of the prediction module and the prefetch module.

3.1. Prediction module
The semantic cache entry of the Web data integration system is defined according to the query characteristics of the Web database, and is described as \( \{ KC, POV, T, W \} \), where \( KC \) represents the set of query keywords submitted by the user, and \( POV \) represents the query period of the next access cycle. The predicted value of the access probability, \( T \) indicates the validity period of the query statement, and \( W \) indicates whether the query statement is prefetched. The basic idea of the prediction model is to calculate the access probability of each semantic cache item in each period according to the unit time as a statistical period. The system collects the access probability of each semantic cache item in the last N periods, based on this. A polynomial regression model that predicts the frequency of
access to the semantic cache entry in the upcoming next cycle. Let the number of total accesses of semantic cache entries in each access cycle be Count (initial value is 0), and the number of accesses per semantic cache entry is Counti (initial value is 0), then each semantic cache in each access cycle The access probability of the item is Pi = Counti / Count. Thereby achieving the real-time prediction. Since the polynomial regression model is linear in terms of parameters, in order to achieve satisfactory results, approximation can be performed by adding high-order measured points. Polynomial can be used to approximate any function in sections, so regardless of the relationship between the dependent variable and other independent variables, polynomial regression can be used for analysis.

3.2. prefetch module

The prefetch architecture includes modules such as query management, storage management, prefetching, cache replacement, and cache coherency [11-13]. The relationship between the modules is shown in Figure 2. Prefetch queues are generated according to conditions such as an access probability prediction value, a prefetch threshold, a prefetch flag bit, and a cache coherency effective time of each semantic cache entry, and prefetching is performed according to the prefetch queue, and the prefetch result set is saved in the outer cache. In the valid data store, the semantic cache description items in the prefetch queue are stored in the effective semantic cache of the inner cache. When each new period comes, a prefetch queue is generated according to whether the relationship between the POV and the threshold α, the data storage identifier W, and the data validity period identifier T are within the validity period to determine which semantic cache items need to be prefetched. Because the prefetching technology has certain speculativeness and the application is not very stable, if the prefetched pages are rarely accessed by users, prefetching increases the useless network services, resulting in waste of network resources [14-15]. The purpose of prefetch control is to expect only prefetching files with a small number but high prediction accuracy, thereby reducing the waste of network resources, shortening the access delay time, and improving the prefetch hit rate. The criterion for measuring the quality of a prefetch algorithm is to see if the algorithm can find a balance between reducing access latency and reducing network bandwidth waste. Therefore, prefetch control is very important for prefetching. The prefetch control can be implemented by setting a prefetch threshold or limiting the maximum bandwidth occupied by prefetching [16]. The following two common prefetch control algorithms are introduced, namely the threshold algorithm and the transmission rate control algorithm.

4. Performance test analysis

The test of prefetch performance is mainly tested from two aspects: query response time and prefetch accuracy. The test case is to access the three test questions from the client, the client user is responsible for querying the access, submitting the query form; the proxy server is responsible for completing the query submitted by the user, performing statistical prediction on each access record and generating a prefetch queue, completing each Prefetch activity for the cycle (24 hours). Generate user query records in ten statistical periods. The first five statistical periods are used as prediction initial data, and the access probability of each query record in each access period is counted, and then the access probability values of each record in the next access period are performed, prediction.
4.1. prefetch module

The user query response time obtained after the prefetching system is not added to the web data integration system and added to the prefetching system is compared and analyzed. The pre-fetch system was not added at the beginning of the test. When the inner cache is 0 semantic cache entries, the average response time of the query is recorded. After adding the prefetching system, when the cache stores 100, 200, 300, and 400 semantic cache entries respectively, the average response time of the query is queried in the sixth, seventh, eighth, ninth, and tenth cycles, and the statistical graph is as shown in Figure 3. Analysis of Figure 3 shows that when the prefetching system is not added to the Web data integration system, the query response time of the user is between 80.35 and 85.25ms, and the average query response time is the highest in each cycle. After adding the prefetching system to the Web data integration system, it can be seen from the comparison of experimental data that with the increase of the number of semantic cache items in the memory, the average response time of each cycle query is steadily reduced, which significantly improves the query response speed of users.

4.2. Prefetch accuracy test

The internal cache stores 100, 200, 300, 400, and 500 semantic cache items respectively, and takes the sixth, seventh, eighth, ninth, and tenth cycles for prefetch accuracy analysis. The test data is shown in Figure 4. Analysis of Figure 4 shows that in each cycle, as the number of caches increases, the prefetch accuracy rate increases. In the sixth cycle, when the number of caches reaches 500 from 100, the prefetch rates are 38.30%, 43.77%, 49.21%, 59.60%, and 61.53%, respectively, showing a steady upward trend. When the number of buffers is 400, the sixth, seventh, eighth, ninth, and tenth cycle prefetch rates are 59.60%, 61.75%, 54.14%, 51.90%, and 59.90%, respectively. Analysis of the data shows that the accuracy of the seventh cycle prefetch is up to 61.75%. Therefore, when selecting the number of caches, a relative balance between the number of caches and the accuracy is also found, so that a higher prefetch accuracy can be obtained by using a smaller number of caches.

5. Conclusions

In this paper, the polynomial regression pre-fetching model is preliminarily studied. By analyzing the prefetching technique, semantic pre-fetching is adopted in the Web, and the predictive value algorithm for predicting the access probability of each semantic cache item in the prediction module adopts polynomial regression technique. After the prefetch queue is generated, the proxy server prefetches according to the prefetch queue. Through the test and analysis of the Web query data added to the polynomial regression prefetch model, the exactness and stability of the polynomial regression prefetch model are proved, which can reduce the network delay and improve the query response speed. Thereby it reducing certain network traffic and reducing network pressure.
Acknowledgement
It is a project supported by Jiamusi University Science and Technology Surface Project (L2013-078).

References
[1] Hong Wei, Chen Meng. :Journal of Computer Products and Circulation, 2018(09): 33+117.
[2] Tian Yugen. Research on Web-based caching and prefetching integration technology [D]. Yanshan University, 2010.
[3] Mukund Deshpande, George Karypis.: Journal of ACM Transactions on Internet Technology (TOIT), 2004, 4(2), 11-14.
[4] Yang Chungui, Wu Zaile, Peng Hongyan : Journal of Computer Engineering, 2007(03): 43-44+47.
[5] Zhou Yangfa. Research on caching technology of Web proxy server [D]. Beijing University of Posts and Telecommunications, 2014.
[6] Yao Yao. A new method of Web prefetching based on client-server double-ended deduplication [J/OL]. Computer Technology and Development, 2019(04): 1-5 [2018-12-20], http://kns.cnki.net/kcms/detail/61.1450.TP.20181220.1035.034.html
[7] Wei Chen, Zhao xiong, Huang Heyan: Journal of Mini-micro Systems, 2004(05): 836-842.
[8] Zhong Yanqing. Research on performance optimization based on Web caching and prefetching technology [D]. Jiangxi University of Technology, 2014.
[9] Sulaiman, Sarina, Shamsuddin etc. : Journal of EN, 2011, 21(5): 88-91
[10] Bin Liu, HaizhenShi, LeiShi etc. : Journal of Software, 2011, 6(9): 24-26
[11] Mei Gaoyong. Research on Web Cache Technology [D]. Northeastern University, 2009.
[12] Ma Hongyuan, Wang Bin: Journal of Information. 2011, 25(5): 37-43.
[13] Yang Chungui, Wu Zaile, Peng Hongyan. etc.: Journal of Computer Engineering, 2007(03): 43-44+47.
[14] ZHAO Zheng, ZHANG Gang, YANG Jie etc.: Journal of Tianjin University, 2001(05): 563-567.
[15] LI Guoqiang, LI Jiashan. : Journal of Computer Engineering, 2009, 35(15): 76-78.
[16] B. Wu, A.D, Kshemkalyani. Objective-optimal algorithms for long-term Web prefetching. IEEE Transactions on Computers. 2006, 55(1): 2-17.