Weizhe ZHANG[1], Hongli ZHANG[1], Hong ZHANG[2], Gui CHEN[1], Yifan WEI[1]

School of Computer Science and Technology, Harbin Institute of Technology, China (1)
National Computer Network Emergency, Response Technical Team/Coordination Center of China (2)

A web partition algorithm based on support vector machine

***Abstract**. In order to solve the problem of network traffic partition of crawler nodes and website nodes, a support vector machine web partition algorithm has been proposed. The algorithm reduces the load of the information collection system on the network through feature extraction, feature selection and support vector machine optimisation, thereby enhancing the response rate and crawling rate of the crawlers.*

***Streszczenie.** Zaproponowano algorytm VSM do rozwiązywania problem podziału sieci ruchu na partycje przez węzły typu crawler i website. Algorytm redukuje przeciążenie napływem informacji przez ekstrakcję cech, selekcję cech i optymalizację VSM. (**Algorytm partycjonowania sieci basujący na maszynie VSM**)*

Keywords: network traffic, web partition, SVM.
Słowa kluczowe: sieci, partycjonaowanie sieci, VSM

Introduction

Crawlers play important roles in the information collection system. The crawler nodes distributed on the WAN capture different websites and have huge discrepancy in the speed and response time. Those crawlers with shorter network distance can have shorter communication latency to the website, as well as a faster webpage downloading speed [1-4]. How to dispatch the captured website to the crawler with a closer network distance, and how to avoid a large scale calculation of network distance are important issues. Inheriting the network distance prediction technology under the application layer, using the predicted distance to classify web space, putting the websites and crawler nodes with closer network distance to a similar set, are therefore all of important research value.

In order to tackle the web partition problem, first this paper formulates the web partition problem. Second, a Web partition algorithm based on support vector machine is proposed. Finally, the performance is verified extensively

Web partition problem

Let all the pages on the web are in the set *W*, $\beta_i\left(i=1,2,...,N\right)$ be subsets of *W*, and *B* be the set formed by β_i . To set $B=\{\beta_1,\beta_2,\beta_3,...,\beta_N\}$ if $\beta_1\cup\beta_2\cup...\cup\beta_N=W$ and $|\beta_i\cap\beta_j|<\delta\left(i,j=1,2,...,N;i\neq j\right)$ (δ is a very small integer,it means intersections between subsets should be as few as possible) , then the nodes $\beta_i\left(i=1,2,...,N\right)$ under set *B* is called the web partition set of the website. The process of dividing the Web is called Web set partition.

(1) Defects of the former web partition algorithm

First of all, although the Chainsaw algorithm can balance the system load, it does not take into account the elements of the locations of the nodes, resulting in a huge consumption of network distance. Secondly, HONet is suitable for short-distanced within-class, and long-distanced inter-class. Otherwise the outcome of partition is linked to the order of sample selection. This will usually lead to too much samples in the first partition set. Moreover, IWAP algorithm cannot ensure every partition set to have or have sufficient crawler nodes. Thus the websites in the crawler-deficient partition set do not have the corresponding crawler capturing, resulting in a scheduling failure.

(2)Advantages of SVM algorithm

First of all, SVM has strong generalisation ability. It is very fault-tolerant and good at partition when handling data with much background noise. The accuracy of the coordination constructed from network coordination system is around 80%. This means the strong generalisation ability of SVM can do a good job in tolerating the discrepancy resulted in the calculation of coordination [5, 6]. Secondly, SVM can solve non-linear partition problem, by projecting non-linear partition interface from low-dimension to high-dimension. Since the system needs to dispatch a huge amount of websites to different crawlers, more than one partition interface is needed. This need is met by SVM. Secondly, SVM has a fixed partition number which avoids the extreme situation of no crawler capture in a website, and it also overcomes the uncertainty of the number of partition sets by using partition algorithm instead of clustering algorithm. To sum up, to use support vector machine in web partition is a feasible method and meaningful try [7-13].

Web partition algorithm based on support vector machine

By the above mentioned, we can have a general idea of the overall process of the web partition algorithm based on support vector machine. The detailed process is shown in algorithm 1.

```
Algorithm 1:  Web partition algorithm

Input: Coordination information of all websites waiting to be classified:  Struct $X_i$ , $i$=1,2…1000
Output: Partition information of websites classified by LibSVM: $Y_i$ ,$i$=1, 2…1000
Statistics information: Statics_LibSVM

Begin
while(partitionresult<predicted result_sat)
  {
      //Use the method of Leave One Out to do Web partition experiment
    Set LibSVM kernel-type param→ svm_type

    param→ kernel_type
    Set LibSVM kernel punishment degrees C and other parameters
    Divide all websites into $N$ groups//Let $N$=10
    for ( $k$=1…$N$)group
  {
Choose 1~k-1,k+1~NgroupWebsite samples to initialize the LibSVM training
{
prob→l:Number of training samples in Web site
Initialization of feature vector array;
prob→y:Labels of training sample categories;
```

```
    Initialize the array according to the selected sample of Categories
    prob→x:Number of training samples;
    }
    Call SVM_ train () to do LibSVM training {
    SVM_Model determined

    SVM_Kernel_Type determined
    Call SVM_train_one () to do LibSVM training
    Received the support vector after LibSVM partition
                                                  }
    // storage of Trained SVM model;
    SVM_Save_Module(k);
    // Select the K-th group of samples of Web nodes to test the partition of Web
    SVM_Web_Partition_Predict(k);
    // Test results are saved into Web
   }//end for
    }//end for
    Obtain the result of Web partition: result;
    Obtain the statistical result of Web partition: Statics_LibSVM
 End
```

Experiment of Web partition

In order to verify the performance of LibSVM partition algorithm, the experiment compares the classifying results of LibSVM from two aspects: one is to use assemble radius clustering criterion function Je to measure the coupling degree of set partition; the other is to apply the classified results to distributed information collection system and scheduling system, and to measure the quality of partition by using network distance consumption [14-16].

The experiment uses Topology models One-level Waxman, Top-down Waxman-Waxman, Top-down Waxman-Barabasi&Albert, and partition algorithm TopK, Chainsaw, HONet, Binning, IWAP. It compares five non-optimal partition methods in the assemble radius and Je value comparison stage [15-18].

Comparison result of collection radius and average Je value

Figure 1 shows the comparison results of the collection radius of five partition algorithms under three network topology models directly. The blue curve represents the trend of the One-level Waxman topology model, red curve represents the Down Waxman-Waxman model, and green curve is Top-Down Waxman-Barabasi&Albert model.

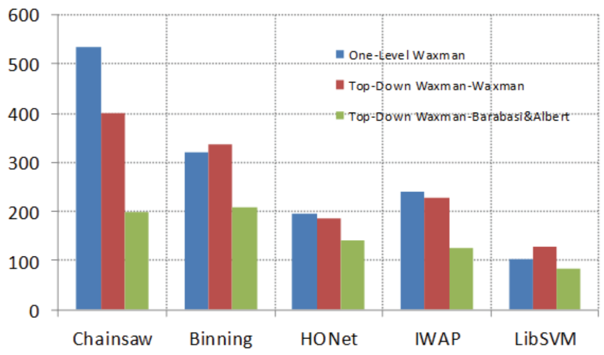

Fig.1. The comparison results of assemble radius

As shown in Figure 1, LibSVM has apparent advantage in assemble radius. In all three topology model, LibSVM's assemble radius is 19.1%、32.4%、42.6% of Chainsaw, 31.8%、38.5、40.6 of Binning, 52.2%、69.7、58.7 of HONet and 64.5%、56.8%、66.8% of IWAP respectively. That is to say LibSVM has obvious advantage in assemble radius, which is due to two reasons: one is LibSVM partition algorithm has fixed number of set numbers, avoiding the situation where there is no crawlers in a specific class based on distance, which needs cross region dispatch of crawlers; the other is fixed partition number (same as the number of crawler nodes) leads to many different kinds of partition, resulting in shorter radius. Overall, LibSVM has certainly advantages with assemble radius.

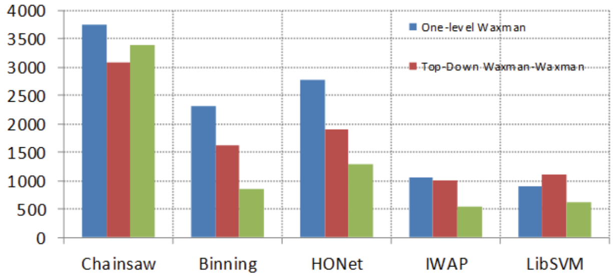

Fig. 2 The comparison results of Je Value

Fig 2 shows the comparison result of 5 partition methods in all three network topology models. The blue curve is the trend of One-level Waxman model, red curve is of Waxman-Waxman model, green curve is of Top-Down Waxman-Barabasi&Albert model.

As shown in Figure 2, in all three topology models, Chainsaw partition algorithm has a Je value much higher than others. The overall performance of LibSVM is relatively outstanding. In all three topology models, the Je value of LibSVM is 39.16%、67.25、72.06 of Binning, 32.52%、56.44、47.84 of HONet and 86.18%、109.01%、113.02% of IWAP. It can be seen that apart from IWAP algorithm, LibSVM is obviously better than others. When comparing to IWAP, LibSVM is slightly worse in Waxman-Waxman and Top-down Waxman-Barabasi&Albert models. This is due to with the topology model, as well as the fact that IWAP is a clustering algorithm which is better at controlling Je value. The author would improve LibSVM in future work to lower the Je value, enhancing the performance.

Comparison of information collection system network distance total consumption

Figure 3 shows the comparison of the total network distance of crawler system in a distributed information collection system under three different topology models. The blue curve represents the trend of the One-level Waxman model, red curve of Top-Down Waxman-Waxman model and green curve of Top-Down Waxman-Barabasi&Albert model.

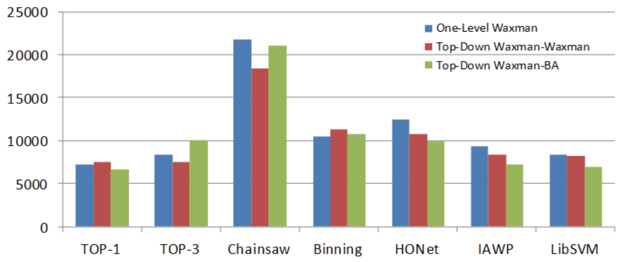

Fig. 3 Network distance total consumption comparison

As shown in Figure 3, performance of LibSVM total network distance is slightly lower than that of TOP-1 classifying algorithm, slightly higher than IWAP algorithm and better than Chainsaw, Binning and HONet. In all three topology models, the total network distance of LibSVM is

116.5%、110.2%、104.3% of TOP-1 respectively and consumes an average of 10.3% more network distance than TOP-1; 98.7%、109.6%、79.4% of TOP-3 and consumes an average 5.1% less of network distance; 79.6%、72.7%、64.8% of Binning and consumes an average 26.4% less of network distance; 67.5%、76.1%、70.3% of HONet and consumes an average 28.7% less of network distance; 90.6%、89.2%、91.3% of IWAP and consumes an average 9.6% less of network distance. By the aforementioned, since TOP-1 and TOP-3 algorithms need to calculate every time in real system, they are not practical to predict distance and can only be used in comparison. Chainsaw algorithm is much worse and therefore there is no need to put it into comparison. LibSVM is better than all other algorithms in total network distance consumption, with a least increase of 9.6%. It has an obvious reduction in network distance consumption in information collection system, as well as the load on network, thereby increasing the response rate and download rate of crawlers.

Conclusion

By introduce the concepts and theories of Web partition, the author has a close look at several typical web partition algorithms, and proposes a web partition strategy based on network distance prediction technology and support vector machine. It classifies the crawler nodes and website nodes in the information collection system of the web space, and compares the classifying result s with other methods.

In the experiment section, we use the support vector machine to classify the web under three typical topology models. We classifies the web effectively through feature extraction, selection and continued improvement of LibSVM. The expected outcome of lowering the network distance consumption of the information collection system is achieved, and the load of the information collection system on the network is reduced. All these increase the response rate and crawling rate of crawlers and pave the way for better performance of the distributed information collection system on the WAN.

Acknowledgments: This work was supported in part by NSFC under Grant Nos. 61173145, National High Technology Research and Development Program of China under Grant No. 2010AA012504 and 2011AA010705, National Basic Research Program of China under Grant No.G2011CB302605.

REFERENCES

[1] Brin S, Page L. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*. 1998,30(1--7):107-117

[2] Barroso L, Dean J, Hoelzle U. Web Search for a Planet: The Google Cluster Architecture. *IEEE Micro*. 2003.

[3] Cambazoglu B B, Karaca E, Kucukyilmaz T, et al. Architecture of a grid-enabled Web search engine. *Information Processing and Management*. 2007, 43(3): 609-623

[4] Baeza-Yates R, Castillo C, Junqueira F, et al. Challenges in Distributed Information Retrieval. In: *International Conference on Data Engineering (ICDE)*.Istanbul, Turkey: IEEE CS Press, 2007.

[5] Heydon A, Najork M. Mercator: A Scalable, Extensible Web Crawler. *World Wide Web*. 1999, 2(4): 219-229.

[6] Boldi P, Codenotti B, Santini M, et al. Ubicrawler: A scalable fully distributed web crawler. In: *The Eighth Australian World Wide Web Conference (AUSWEB02)*.2002.

[7] Chang F, Dean J, Ghemawat S, et al. Bigtable: A Distributed Storage System for Structured Data. In: *OSDI'06: Seventh Symposium on Operating System Design and Implementation*.2006.

[8] Karger D, Sherman A, Berkheimer A, et al. Web caching with consistent hashing. *Computer Networks*. 1999, 31(11-16): 1203-1213.

[9] Soderland S. Learning Information Extraction Rules for Semi-Structured and Free Text. *Mach. Learn*. 1999, 34(1-3): 233-272.

[10] N. R. Sakthivel, V. Sugumaran, Binoy B. Nair Application of Support Vector Machine (SVM) and Proximal Support Vector Machine (PSVM) for fault classification of monoblock centrifugal pump. Dec. 2009:38- 61

[11] Erik Boiy, Marie-Francine Moens A machine learning approach to sentiment analysis in multilingual Web texts. Oct. 2009 *Information Retrieval* : 526 – 558

[12]T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications to pattern recognition. *IEEE Transactions on Electronic Computer*s, 1965, 14(3): 326-334

[13] T S Eugene Ng H Z. Predicting Internet Network Distance with Coordinates-Based Approaches. New York: *Proc of IEEE INFOCOM*, 2002.

[14] Guyton J D, Schwartz M F. Locating nearby copies of replicated Internet servers. *SIGCOMM Comput. Commun. Rev*. 1995, 25(4): 288-298.

[15] T S Eugene Ng H Z. Towards Global Network Positioning. In: *ACM SIGCOMM Internet Measurement Workshop*.San Francisco, CA: 2001.

[16] Ratnasamy S, Francis P, Handley M, et al. A scalable content-addressable network. *SIGCOMM Comput. Commun. Rev*. 2001, 31(4): 161-172.

[17] Paul Francis S J C J. IDMaps: A Global Internet Host Distance Estimation Service. *IEEE/ACM Transactions on Networking*. 2001, 9(5): 525-540.

[18] Yager R R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern*. 1988, 18(1): 183-190.

***Authors**: prof. Weizhe Zhang, Box 320, School of Computer Science of Technology, Harbin Institute of Technology, Heillong Jiang Province, P.R.China,150001 E-mail: wzzhang@hit.edu.cn;*