RESEARCH ON AN ANTI-CRAWLING MECHANISM AND KEY ALGORITHM BASED ON SLIDING TIME WINDOW

Yi Liu , Zhengqiu Yang , Jiapeng Xiu , Chen Liu

Beijing University of Posts and Telecommunications, Beijing 100876 , China
liuyi@goldensystem.cn

Abstract: Inadequate crawling behavior of the crawler will have a very serious impact on the site, so anti crawling mechanism is an important function for the website. Most of the existing anti crawling methods are non real time detection, and the recognition accuracy is low. By analyzing the characteristics of Crawler，a real-time crawler detection method based on sliding time window is proposed, which improves the accuracy and efficiency of detection of non compliance with the rules of Crawler.

Keywords: Anti-crawling; Sliding time window; Crawler asynchronous detection;

1 Introduction

With the development of search engine, web crawlers have become a very popular network technology. Almost every large portal site has its own search engine, such as Google, Yahoo, Baidu and other well-known search engine. There are nearly one hundred kinds of known search engine crawlers, let alone the unknown ones, which are uncountable. For a content driven web site, it is inevitable to be visited by web crawlers. The disadvantage of this situation is that a large number of bandwidths are occupied by crawler.

Some intelligent search engine spiders crawling frequency is reasonable and it consumes less network resources, but there are many unreasonable designs of web crawler that web crawling ability is very poor and often complicated by dozens of hundreds of repeated requests to crawl .It is a devastating blow to the small and medium sites., especially some crawlers which were written by programmers who lack of experience. It would cause access slowdown and service disruption. Therefor, it is a common practice of web sites to build up an anti-crawl mechanism. The basic principles of anti crawling are recognizing the behavior of crawlers and rejecting crawlers' access. There are several methods to recognize crawlers currently. The first way is to analyze the features in different access behavior and then adopt some method like threshold, decision tree and Bias network [1] to check web crawlers. However, one disadvantage of this method is its hysteresis and time-consuming for it can't find and dispose the crawl request by processing in real-time. The second way is based on trap technology[2].You can set a hidden link on the web page which can't click by users but can be visited by web crawlers. We can trap crawlers accurately and timely in this way, while it would be useless when designers realize the trap. The third way is based on user access request[3].It is able to judge the crawler's characteristics timely but not accurate enough, for every access request needs to be checked, it would affect the user experience if verification logic too complicated. Based on all above, one of the most important research currently is to build up a more efficient and accurate anti-crawling system.

A new mechanism of anti-crawler is proposed basis on analyzing the characteristics of crawler access and sliding time window in this paper, and the key algorithms are studied also. This mechanism can make fast and accurate to identify crawler and use less resource.

2 Anti-crawling mechanism based on sliding time window

2.1 Comparative and analysis of the behavior of users and crawls

Due to their natural attributes and purposes of the request are different, web requests from users or crawlers show different behavior patterns. The differences in access are listed in Table I.

Table I Differences between user requests and crawler requests

Differences in access time		
	user	crawler
Access rate	Rate has a ceiling	Fast and stable
Access Interval	Irregular	Orderly
Number of requests per unit time	Number has a ceiling	as many as they can
Differences in access content		
The proportion of pictures and words in the request	Proportional balance	topical crawler are disproportionate
Judgment cookie	normal	The value of cookie is null, or can't pass verification
Judgment session	normal	Large number of requests in a single session
Judgment referer	normal	can't pass verification

978-1-5090-1256-5/16/$31.00 ©2016 IEEE

As shown in the table I, the differences of behavior characteristics between users and web crawler are in two aspects: one is different in access time. Users read web pages sometimes fast and sometimes slow because they are affected by the surrounding environment and physiological factors when they are reading .The other one is different in access content. The reason is that users who read a web page with a less sense of purpose relative to crawler which get data.

So we study from these two aspects, subdivision and set weight. We can distinguish between crawler and users base on our research.

2.2 Design of anti-crawlers mechanism

One of the demand of this system is that there’s no need to change a lot on the base of original system, when the already running web project needs to add this system. So the main objective of this system is that it is high cohesion and low coupling between original system and anti-crawlers system. In order to achieve this objective , the system consists of two components- anti-crawler client and anti-crawler server. Each component has several processing modules as shown in the following figure.

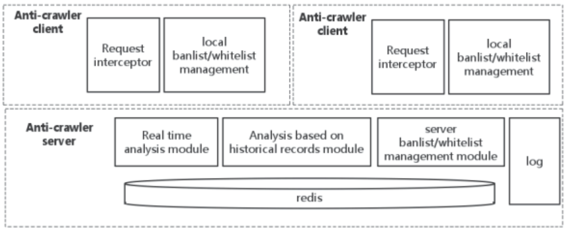
Anti-crawler client
Request interceptor
local banlist/whitelist management
Anti-crawler client
Request interceptor
local banlist/whitelist management
Anti-crawler server
Real time analysis module
Analysis based on historical records module
server banlist/whitelist management module
log
redis

Figure 1 Anti-crawler system functional diagram

The client component consists of request interceptor module and local banlist/ whitelist management module. The server component consists of real-time analysis module , analysis based on historical records module, server banlist/whitelist management module, log management and data module. The sequence diagram of the system as shown in the following figure.

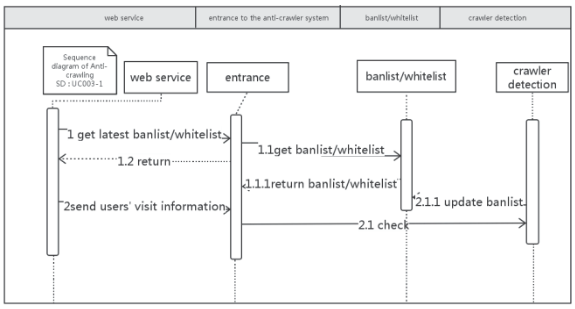
web service
entrance to the anti-crawler system
banlist/whitelist
crawler detection
Sequence diagram of Anti-crawling SD : UC003-1
web service
entrance
banlist/whitelist
crawler detection
1 get latest banlist/whitelist
1.1get banlist/whitelist
1.2 return
1.1.1return banlist/whitelist
2.1.1 update banlist
2send users' visit information
2.1 check

Figure 2 Sequence diagram of Anti-crawling

As shown in the figure 2,it's necessary to add the following functions in anti-crawler client component .

Function 1: judge whether users’ access in the banlist/ whitelist.

Function 2: send users' visit information to Anti-crawler system.

Function 3: get the latest banlist/ whitelist.

The following functions must be added in anti-crawler server component .

Function 1: accept the client requests and judge type of requests.

Function 2: CRUD operations of banlist/ whitelist.

Function 3: Detection-Detecting malicious activity of crawlers.

Function 4: send the latest banlist/ whitelist to client .

Based on all above, we could know that web service realize an anti-crawler system without consuming too much web service system resources. This is principally because that the anti-crawling which is made by anti-crawler system is not in original web service, so that it cause minor effects of users' visit as much as possible.

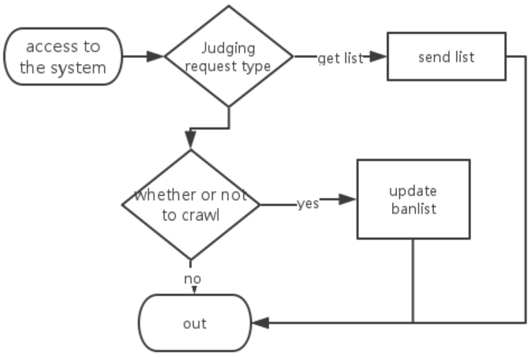

Figure 3 Operation flow chart of the anti-crawler system

From figure 3, we conclude that the core part of anti-crawlers system is the detection way of crawler, the method we use is SVM which based on sliding time window. It will be discussed in more detail at a later stage.

3 Anti crawling mechanism design

3.1 concept of sliding time window

Sliding time window is a common method to deal with time series. At first, introduce the related concepts of Sliding time window.

Definition 1. Time window

Given a event set of user access E and a user access sequence $S_w = \{s, T_s, T_e\}$, *Ts* is the sequence of events beginning user access time. *Te* is the sequence of events the end user access time.

s is a sequence of increasing user access, namely

$$s = \{(e_1, t_1), (e_2, t_2) \dots \dots (e_n, t_n)\}$$
$$e_i \in E, T_s \leq t_i \leq T_e, i = 1, \dots \dots ,\ n$$

Subsequence

$$S_w = \{w, t_s, t_e\}$$

called a time window of S, among them,

$$T_s \leq t_s \leq t_e \leq T_e, w = \{w \subseteq \mathrm{S} | t_s \leq t \leq t_e\}$$

among them,$t_e - t_s$,Called window width，as W。

Definition 2. Slide operations .

There are two operations, move and continuous query for the time window which used to store session information. Restrictions by system memory size and the amount of existing users, We need to control the amount of data in each time window. When a new access of a user's request comes we need sliding window to keep user's recent web access in a window, called the slide operation.

Definition 3. The sliding window step size *S*.

Suppose that a sequence of user access (Referred to as the access sequence) in the current window start time is T_i and end time is T_e and T_e -T_i =W. The new start times is $Ti + s$ and new end times is $Te + s$ when the window slide a step S

Sort of things is a more common applications, including e-mail, text classification, Member, behavior, safety, and so do Office classification. The current understanding of the situation: Bayesian, decision tree, neural network, SVM is a popular several classification algorithms. SVM is the algorithms selected by this paper to used to categorize

Definition 4. SVM algorithm

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

The advantages of SVM[4]:

Recent algorithms for finding the SVM classifier include sub-gradient descent and coordinate descent. Both techniques have proven to offer significant advantages over the traditional approach when dealing with large, sparse datasets—sub-gradient methods are especially efficient when there are many training examples, and coordinate descent when the dimension of the feature space is high.

3.2 SVM based on sliding time window

Determining the step of sliding window

A suitable step is more important for this algorithm. The step of sliding in this paper is not fixed. The initial value of the step size is 27 and the maximum value is 40.We can determine whether the newly added data can be placed in a window with the existing data according to the following three conditions when the number of data in the window more than 27. Is this web access and the last web access in the same session? Is the interval between this web access and the last web access is smaller than the average interval in the window? Will the referer of this access join to the referer forest of sliding window change the number of forest? We make the step plus 1 if there are two or more yes .We must check if the window step is more than 40. the window step size as a dimension can be judged.

Determine the dimension of feature space

1: Window step size can be used as a dimension.

2: Cookie can be used as a dimension. There are three states of cookie .Cookie does not exist, cookie exists but the check fails, cookie exists and check success. Number of cookies that are not repeated in a sliding window can be used to judge.

3: Session can be used as a dimension. I was mainly judge from the perspective of three: the number of sessions that are not repeated in a sliding window, whether a single session is corresponding to different IP, the maximum number of session window access

4: Access time can be used as a dimension. I was mainly judge from the perspective of three: maximum margin of time, the median time between accesses, intervals for variance of time.

5: Referer can be used as a dimension. I was mainly judge from the perspective of three: referer does not exist, referer exists but isn't our site's, referer exists and is our site's.

6:User agent can be used as a dimension. I was mainly judge from the perspective of four: there is no declaration, declare as an unknown crawler, declare as a known crawler but the IP is not correct, declare as a known crawler and IP is correct.

When these 6 dimensions of a access are determined, unit of it and use as input.

3.3 SVM

The main goal of this model is determine the maximize distance of user access. In the end, the problem that detection crawlers is transformed into a convex quadratic programming problem. The most important of the model is to obtain the maximum geometric distance of the event of user access .Makes all points of the interval greater than the geometric interval.

Step1: Given an set of user access event

$$D = \{(X_i, y_i) | X_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n}$$

In this formula, *xi* is the feature vector, *yi* is the difference between crawlers or users, *p* is the dimension of feature vector, *n* is the number of samples.

Step2: Using a linear classifier put these data into two categories. The learning goal of linear classifier is to find a hyper plane in n-dimensional data space，This hyper plane equation can be expressed as:

$$xw^t + b = 0$$

*Step3:*After given a training set and a super plane，Determined interval function by

$$\gamma\hat{} = y(xw^t + b) = yf(x)$$

and obtain the minimum function interval

$$\gamma\hat{}= min\,\gamma\hat{}i \ \ (i=1， ...n)$$

and geometric interval

$$\gamma = \frac{(xw^t+b)}{||w||} = \frac{f(x)}{||w||}.$$

The objective function of the maximum distance classifier can be defined as:$Max\ \gamma\hat{}$.We need to meet some conditions according to the definition of interval,

$$y_i(x_i w^t + b) = \gamma\hat{}_i \geq \gamma\hat{}_i, i = 1, \ldots, n$$

Step4: Make the function interval $\gamma\hat{}$ equal to 1,so $\gamma\tilde{}$ = 1 / ||w|| and

$$y_i(x_i w^t + b) = \gamma\hat{}_i \geq 1, i = 1, \ldots, n$$

Thus the objective function is transformed into

$$max \frac{1}{||w||}, s.t, y_i(x_i w^t + b) = \gamma\hat{}_i \geq 1, = 1, \ldots, n$$

Step5: Maximum value of $\frac{1}{||w||}$ can be converted to the minimum required for the sake of $\frac{1}{2}\left||w|\right|^2$.So the above objective function is equivalent to:

$$min \frac{1}{2}\left||w|\right|^2, s.t, y_i(x_i w^t + b) = \gamma\hat{}_i \geq 1, = 1, \ldots, n$$

Step6: Because the objective function is quadratic, the constraints are linear, so it is a convex quadratic programming problem, and need to verify the conditions *kkt*:

$$min. f(x)$$

$$s.t.\ h_j(x) = 0, j = 1, \ldots, p,$$

$$g_k(x) = 0, k = 1, \ldots, q\ x \in X \subset R^n$$

S*tep7 :*We can get solution of quadratic programming by verification the solution .value of w and value of b in the $f(x) = (xw^t + b)$ can be obtained.，We can get the maximum interval separating hyper plane

$w^T + b = 0$ and a classification decision function

$$f(x) = sign(W^t * x + b)。$$

Thus, the algorithm model has been established and the new user access request can be classified and judged.

4 Conclusions

We put forward a real-time anti-crawler algorithm based on sliding time window and classification algorithm after researching of the characteristics of crawlers and summarizing the traditional ways to detect crawlers. The advantage of this system is that it improves accuracy rate，while lowers the cost of system resource consumption and reduces a amount of unnecessary operations. We make amount of data as an input, it would improve the dimensions, and reduced the computation times. Our direction of research will mainly concentrate on these three aspects. Using this system in large scope if conditions permit, combining trap technology with our system to improve efficiency and success rate of detection and determining the survival time of every object in banlist.

References

[1] Guo Weigang, take time. [J]. detection technology of Computer Engineering in e-commerce website Web Robot,2005,31(23):219-221.DOI:10.3969/j.issn.1000-3428.2005.23.081.

[2] Fan Chunlong, Yuan Bin, Yu Zhou Hua et al. A network crawler based on trap technique for detecting [J]. computer applications,2010,30(7):1782-1784,1793.

[3] A survey on the detection technology of Wu Xiaohui, Ji Xing.Web crawler [J]. Journal of Hubei Automotive Industrial Institute, 2012,26 (1): 57-59,72.DOI:10.3

[4] Song ting. Research and implementation of web crawler detection based on [D]. SVM Tianjin University, 2010.

[5] KLEMETTINEN M, MANNILA H, TOIVONEN H.Rule Discovery in Telecommunication Alarm Data [J]. Journal of Network and Systems Management, 1999, 7(4): 395-423.

[6] HAUPTMANN M, LUBIN J H, ROSENBERG P,et al. The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk [J].Statistics in Medicine, 2000, 19(1): 2185-2194.

[7] P N Tan;V Kmmar Discovery of Web Robots Session Based on their Navigational Patterns2002(01)

[8] DORAN D, GOKHALE S S. Web robot detection techniques: overview and limit ations[J]. Data Mining & Knowledge Discovery, 2011, 22(1-2):183-210.

[9] BOMHARDT C, GAUL W, SCHMIDT-THIEME L. Web Robot Detection - Preprocessing Web Log files for Robot Detection [J]. Studies in Classification Data Analysis & Knowledge Organization, 2004:113-124.

[10] SUYKENS J A K，VANDEWALLE J. Least squares support vector machine classifiers［J］. Neural Network Letters，1999，9 (3) : 293-300