JAIST Repository
https://dspace.jaist.ac.jp/

Title	稀にアクセスされるWebページを検出するシステムに関する研究
Author(s)	立花，一樹
Citation	
Issue Date	2012-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/10504
Rights	
Description	Supervisor:知念賢一特任准教授，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

The method for detection system of infrequently Web pages

Kazuki Tachibana (0810037)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 6, 2012

Keywords: rare Web access, long-term observation, power-law.

Since 1991 when World Wide Web leaded to the invention of Tim Berners-Lee, there are many services which is using HTTP such as electronic payment, internet news, Web-Mail. As a result of increased of their services, the most part of traffic on Internet is using HTTP. As a variety of user is diverse, there are many kinds of the aim which access to Web page. As a result, the amount of information which is accessed entire Web pages is gargantuan. But the relationship between frequence and rank based on frequence accordance with power-law. So, it is widely known that the most part of access is concentrate to frequently Web pages. Today, a part of frequency of appearance URL is found out in detail. But It hardly know the other pages in Web access. In such access, there are posssibility not only just non-popular Web pages but also dangerous Web pages which is managed by malicious user. Therefore, I proposed the system which is able to go live for an extended period of time to detect and record the URL of infrequently Web pages. And if we will find the access pattern to infrequently Web pages, it has important siginification and knowledge at social end. For example, we will be able to detect dangerous access which is not able to detect until it.

Firt of all, I defined the target area in term of the relationship between frequency and access rank base on it. And we design the ovservation system which can detect only the URL of infrequently Web pages in real time for

Copyright © 2012 by Kazuki Tachibana

an extended period of time. The system has two important function that the one is that the system has to be able to record all URL characters which is accessed and the other function has to be able to discriminate only infrequently Web pages efficiently. So, the system has independent data structure which has high efficiency of capacity and will be able to find the infrequently URL as fast as possible. The system has to have the method which can detect in frequently URL independent of observation time in limited computational resource. So, the system has equally-spaced array to record interval time of access and the frequency of appearance about all Web pages. And this system has the method which have an ability to adapt to estimate value about appearance number of URL in observational environment or hardware specification, or accuracy of observational aim.

I managed this detection system which is input the packet dump file in the network of Japon Advanced Institute Scient and technology for 16 days. In this experience, It spent 1 day as learning time. In this learning term, I found the total appearance number of URL is two million and five hundred-thousands. And the the system start to detect in frequently URL after learning term for 16 days. The total appearance number of URL is twenty eight million for 16 days. The assumption to discriminate infrequently URL is that the interval time is longer than 12[hour] and the appearance frequency of the URL is less than one hundred. This rule is experimental assumption and I think to examine carefully about this parameter. And the apperance number of infrequently URL keeps to increase and the number of URL after 16 days decupled compared to it which is first day of observation. From the result of this detection expriment, it revealed that this detection system cannot record many characters of URL in long-term observation, but the detection mechanism is effective to discriminate rare URL. So, I considerd about some resolution to reduce fixing memory region and compressive solution the characters of URL by using Patricia trie.