# Estimating The Web Robot Population

Yang Sun
AOL Research
888 Villa St
Mountain View, CA, USA
yang.sun@corp.aol.com

C. Lee Giles
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, USA
giles@ist.psu.edu

## ABSTRACT

In this research, capture-recapture (CR) models are used to estimate the population of web robots based on web server access logs from different websites. Each robot is considered as an individual randomly surfing the web and each website is considered as a trap that records the visitation of robots. We use maximum likelihood estimator to fit the observation data. Results show that there are 3,860 identifiable robot User-Agent strings and 780,760 IP addresses being used by web robots around the world. We also examine the origination of the named robots by their IP addresses. The results suggest that over 50% of web robot IP addresses are from United States and China.

## Categories and Subject Descriptors

K.1 [**THE COMPUTER INDUSTRY**]: Statistics

## General Terms

Measurement, Experimentation

## Keywords

web robot population

## 1. INTRODUCTION

More and more online services are enabled by web robots acquiring information from web servers. The increasing trend and variety of services make web robots occupying a significant part of the internet bandwidth. The web server access log analysis on $CiteSeer^x$[1] shows that robot generated visits is exceeding regular user generated visits in 2008. The log analysis on another two Chinese websites and many online discussions about robot traffic (e.g., webmaster world[2]) support the findings that web robots become a major traffic contributor on the Web. One fundamental question is: how many web robots are there? More specifically, how many unique web robot names are used? How many IP addresses are used by these robots? And where are these robots coming from?

[1]http://citeseerx.ist.psu.edu
[2]http://www.webmasterworld.com/


*WWW 2010,* April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Estimating the size of a population is always an interesting task. In this research, we use capture-recapture (CR) models to estimate the population of active web robots. CR models have a rather long history in the biometry literature where they have been used to estimate the population sizes of wildlife animals [7, 3, 4, 5]. This method is valuable in the situation where observing all individuals in a population is not possible. Researchers visit a target area and set traps to capture a group of individuals of a kind of animal. These individuals are marked with unique identifications and released back into the environment. Next, the researchers capture another group of individuals from the same population. The population size can be estimated by applying statistical models to the two data sets obtained from the CR experiments. Recently CR models are applied to estimate the size of the web [1, 2] where search engine indexes are considered as capture marks of web pages. The research assumes that each web page has the same probability of being captured and each search engine is an independent source of capturing web pages. Thus, the number of indexed web pages and overlapped web pages in different search engines can be used to fit in the Lincoln-Peterson model and to estimate the size of the web. The population of telephone lines are estimated in [6] where phone records for a certain period are considered as capture marks. Different time periods are considered as different capture occasions. Maximum likelihood estimation (MLE) method [3] is used to fit the data to advanced CR models.

## 2. MODELS

We use MLE method to estimate the population of web robots. Web server access logs from three different sources are analyzed as capture traps for web robots. The experiments of estimating the web robot population have similar settings to the wildlife animal study. The World Wide Web is considered as an open field for web robots. Websites are traps to robots where researchers can "capture" robots and "mark" them by analyzing the access logs. Each web robot corresponds to an individual in the population that may or may not visit a website. In this setting, web server access logs for a period of time is considered as a capture occasion in which web robots are marked and identified by IP addresses and User-Agent strings.

More generally, we can present the CR model using maximum likelihood of observations. Let $t$ be the number of capture occasions (observations), $N$ be the true population size, $n_j$ be the number of individuals captured in the $j^{th}$ capture occasion, $M_{t+1}$ be the number of total unique in-

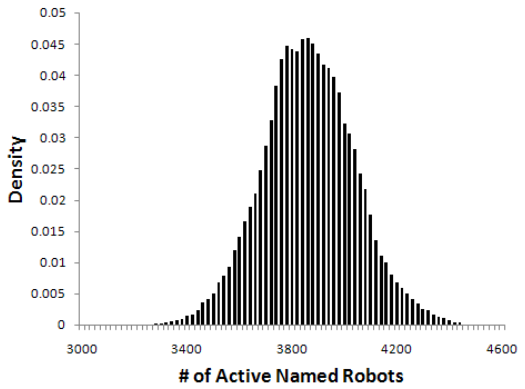


**Figure 1: The estimated web robot name population size $N$ using MCMC method.**

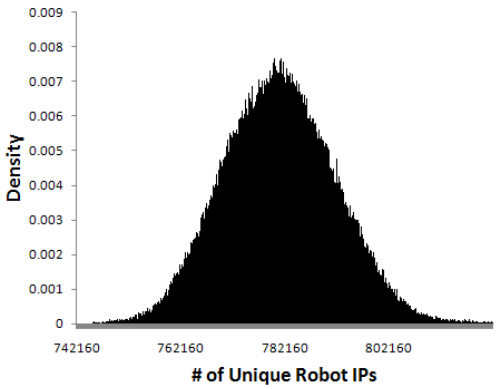


**Figure 2: The estimated web robot IP population size $N$ using MCMC method.**

dividuals caught during all occasions, $p$ be the probability of an individual robot being captured and $f_j$ be the number of robots being observed exactly $j$ times ($j < t$). The likelihood function for the $t$ observations is:

$$L(N,p) = \frac{N!}{(N-M_{t+1})!}\prod_{j=0}^{t}(p^j(1-p)^{t-j})^{f_j} \qquad (1)$$

## 3. RESULTS

We extract data from web server access logs of three websites in May 2008. We assume the three "captures" are independent from each other for the following two reasons: 1. There is no hyperlinks between these websites. 2. The content and geographical location of the three websites are irrelevant to each other. Thus, robots visiting one website will not affect the probability of visiting the other. We use a JAVA MCMC program to obtain samples from the joint posterior distribution described in Equation (1). We ran the Markov Chain for 200,000 iterations with 10,000 burn-in. Figure 1 and 2 show that the identifiable named robot population is about 3,860 and the size of IP addresses used by web robots is about 780,760. On average each robot name uses more than 200 IP addresses. However, since many robots hide their true identity by faking User-Agent strings, the actual number of named robots should be much larger than 3,860.

| Country | Unique IPs | Country | Unique IPs |
|---|---|---|---|
| United States | 426,148 | Germany | 57,758 |
| China | 301,321 | Korea | 38,160 |
| United Kingdom | 108,422 | France | 35,092 |
| Japan | 69,060 | Canada | 36,390 |
| India | 64,414 | Australia | 30,457 |

**Table 1: Top 10 countries have the largest number of robot IPs.**

### 3.1 Geographical Distribution

We also try to study the geographical distribution of web robots. Since the same robot name may include IP addresses from multiple locations, we can only estimate the IP addresses used by robots in different geographical locations. In this research, we segment robot IP addresses by country. The top 10 countries having the largest number of robot IP addresses are listed in Table 1. We only listed the maximum density point for each country instead of the whole distribution for page limitations.

## 4. CONCLUSIONS

The population of web robots is an important fact to the World Wide Web community. In this research, we apply CR models to estimate the population of web robots based on web server access logs. Robots are assumed to have uniform probabilities of visiting a random website. The access logs of three independent websites are analyzed and MCMC sampling method is used to fit the data to CR models and to estimate the robot population. To our knowledge, it is the first research effort to estimate the size of web robot population and total IP addresses used by web robots. Experimental results show that there are 3,860 identifiable robot User-Agent strings and 780,760 IP addresses are used by robots. We also examine the origination of the named robots by their IP addresses. The results suggest that more than 50% of web robot IPs are from United States and China.

## 5. REFERENCES

[1] A. Dobra and S. E. Fienberg. *How large is the World Wide Web?*, chapter Web Dynamics. Springer-Verlag, 2003.

[2] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, 1998.

[3] D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson. Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62:1–135, 1978.

[4] S. Pledger. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56:434–442, 2000.

[5] K. H. Pollock. A capture-recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, 46:757–760, 1982.

[6] D. Poole. Estimating the size of the telephone universe: a bayesian mark-recapture approach. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 659–664, New York, NY, USA, 2004. ACM.

[7] R.King and S. Brooks. On the bayesian analysis of population size. *Biometrika*, 88(2):317–336, 2001.