An Internet real-time programming technique for monitoring cache changes

Ioannis Anagnostopoulos[1] and Christos-Nikolaos Anagnostopoulos[2]

[1]Department of Information and Communications Systems Engineering
University of the Aegean, Karlovassi - 83200, Samos, Greece
janag@aegean.gr
[2]Department of Cultural Technology and Communication
University of the Aegean, Mytilene - 81100, Lesvos, Greece
canag@ct.aegean.gr

Abstract

Due to the extremely rapid growth of the web, Internet services cannot monitor the evolution of its size at the same time or with the same priority. This paper proposes an innovative method for measuring the ability of Internet search services to maintain new and up-to-date results in their directories. The method is based on capture-recapture experiments in real wildlife biological studies. This proposal can be implemented as a real-time Internet service, in order to continuously measure the Internet evolution as well as the capability of Internet search services to maintain new and up-to-date caches.

Keywords: Internet caches, web search, Internet evolution

1. Introduction

As the Internet grows hastily, the increase of web resources has boosted the demand for effective and personalized search among them. Trying to satisfy their information needs, million of users worldwide, submit their queries on heavily-visited Internet search services, which play the role of the information broker between the user and the disseminated information. However, do the search services update their indices and catalogues frequent enough? Is the user aware about how able is a search engine to follow the evolution of the web? After all, it is well-known that the results provided in respect to a user's query, are actually an "image from the past". So, wouldn't be appropriate to endorse the search services, which adequately adapt to the Internet evolution (e.g. index faster the new information, check the validity of their results, remove the dead links etc.) and penalize those that fail to perform such information management (or are not sufficient enough in respect to other services)?

Thus, in this paper we propose real-time Internet-based programming technique, which is capable of monitoring the continuous changes that occur in the caches of major Internet search services.

2. Monitoring evolution with capture-recapture measurements

Our proposal is based on capture-recapture experiments used in wildlife biological studies [1]. In such experiments animals are captured, marked and finally released on several trapping occasions. If a marked animal is captured on a subsequent trapping occasion, it is said to be recaptured. Based on the number of marked animals that are recaptured, using statistical models and their estimators one can estimate the total population size, as well as the birth rate, the death rate and the survival rate of each species. There any many capture-recapture sampling protocols in the literature. The sampling scheme chosen for capturing, marking and recapturing the meta-results is the robust design [2], which extends the Jolly-Seber Model [3]. This model was chosen among other capture-recapture models since in wild-life experiments it is applied to open populations, in which there is possibly death, birth, immigration, and permanent emigration. In this Internet hypothesis, death corresponds to a result, which is no longer exists, while birth match to a new result (new or updated information). The concept of the real-time Internet cache monitoring mechanism is summarized in the following sub-section. In order to adopt the real-life experiments in our algorithm, we considered the assumptions described in [4] and [5]. However, in our case, the third-party results are considered as wild animals, while the population under study consists of the sampled third-party results.

When transferring the experiments on the web, different animal species under study that live in an area correspond to different results derived from different queries, while the traps correspond to our sampling method. In the real life experiment the traps are set up for a specified amount of time and the theory used assumes that all animals have the same probability of being captured in a trap. Thus, in our algorithm we had to ensure that meta-results of different queries have the same probability of being captured during the sampling procedure. Taking into account these considerations the sampling method is as follows. Working in the background the algorithm keeps a file, which records of all the submitted queries of the user. This file is parsed and each record is included to a second sample under a probability value p_1. Then, these queries are simultaneous submitted to the used Internet search services. Then, among a selected amount of results, we randomly select a portion of them respectively, according to a second probability value p_2. Thus, we collect some results in order to mark and examine them as the biologist does in the real-life experiments. Afterwards, for each of the two assembled set we examine whether the results have been appeared in previously samplings. If they are not, then births are occurred. In case where the results are not new we check whether corresponds to an up-to-date

result. If the cached document corresponds to the active disseminated content, then we consider that the examined result is survived, while is not when the cached content is different with the active one. Finally, a result is considered as a dead one in case where its active content is not active. Probability values p_1 and p_2 are fine-tuned according to the amount of the previously submitted queries of the user. Finally, after multiple sampling and taking into account the total amount of captured instances (results) in a current sample, the previously marked as well as the marked instances in subsequent sampling periods, the proposed real-time Internet cache monitoring mechanism estimates the survival rate and the birth rate for the used Internet search services.

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Primary Periods	1st		2nd		3rd		4th		5th		6th		7th		8th	
Secondary Periods	8		16		24		32		40		48		56		64	
Birth (b), Survival (φ) rates		b1 φ1		b2 φ2		b3 φ3		b4 φ4		b5 φ5		b6 φ6		b7 φ7		

Figure 1. *Timetable of a real-time capture-recapture Internet experiment*

3. Capture-recapture measurements

This section presents the results, which were derived after a sixteen-week period of real-time Internet capture-recapture experiments. The time interval between two subsequent primary sampling periods was two weeks, while the respective time for subsequent secondary sampling periods was one day, starting at the beginning of each primary sampling period and conducted for four days. In our case, the secondary periods were performed at the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th day after the launch of the primary one as depicted in Figure 1. The buffer size of the pool consisted of the last 100 user queries, which were used for generating the sampled population. In order to get a measurable amount of tested instances (results), probability values p_1 and p_2 were set equal to 0.5 and 1 respectively. This implies that at each secondary sampling period it was expected to submit approximately fifty queries out of the last 100 last stored in the pool. For every one of the selected queries the top-fifty results from each Internet search service were collected.

3.1 Case Study

Table 1, holds all the respective parameters regarding capture-recapture measurements, which was conducted from September of 2006 to January of 2007 (16 weeks) using two major internet search services, namely ISE1 and ISE2. We intentionally do not cite the Internet search services, since in our objectives was not

the evaluation of the search services used at the moment. Besides, such kind of assessment, which requires repeatedly evaluations before its results are disseminated publicly is left for future work.

Thus, *Ni* stands for the absolute values of the tested results for the used search services, where i=1,…,8 corresponds to the i^{th} primary sampling period. On the other hand B_i, b_i and φ_i, where i=1,…, 7 define the births of new results (in absolute values), the birth rate as well as the survival rate between the i^{th} and the $(i+1)^{th}$ primary sampling periods respectively.

For example, after the completion of the first two primary sampling periods, which consisted of sixteen secondary sampling periods, we investigated that ISE1 included 485 new results in its index (B1) over 23396 total provided results (N1+N2), while ISE2 included 386 new results (B1) over 23326 provided results (N1+N2). Thus, the birth rates between the first and second primary sampling period (between September and October of 2006), were measured at the levels of 4.13% and 3.34% for ISE1 and ISE2 respectively. Having also calculated the refresh rate and the up-to-dateness of these results, the respective survival rates were measured at 96.78% and 94.69%.

Table 1. *Capture-recapture measurements (from september 2006 to january of 2007)*

Param.	ISE1	ISE2	Param.	ISE1	ISE2
N1	11643	11783	**b1**	0.0413	0.0334
N2	11753	11543	**b2**	0.0399	0.0354
N3	11639	11647	**b3**	0.0546	0.0348
N4	11782	11876	**b4**	0.0342	0.0350
N5	11844	11758	**b5**	0.0451	0.0329
N6	11694	11780	**b6**	0.0316	0.0384
N7	11807	11867	**b7**	0.0351	0.0329
N8	11683	11769	**avg(b)**	0.0402	0.0347
B1	485	386	**φ1**	0.9678	0.9469
B2	464	412	**φ2**	0.9508	0.9733
B3	643	413	**φ3**	0.9570	0.9842
B4	405	412	**φ4**	0.9709	0.9554
B5	527	387	**φ5**	0.9428	0.9690
B6	373	456	**φ6**	0.9778	0.9687
B7	410	387	**φ7**	0.9548	0.9591
period: 18/09/06 - 08/01/07			**avg(φ)**	0.9603	0.9652

In other words, for this time interval ISE1 not only indexed more new results but also manage to provide more results that their actual disseminated content was the same with the cached content in its directory during the experiments. After the end of the experiments, the derived results shown that during the period of the experiments (September 2006 to January 2007) the average birth rate for ISE1 was measured at nearly 4%, while the respective rate for ISE2 was reaching 3.5%. In addition, the average survival rate of these two tested Internet search services was also very close and calculated at the levels of 96% and 96.5%. These results confirm that both search

services have virtual equal capabilities in updating their directories and provide new and up-to-date results to their users.

4. Potential applications

In the following we describe some potential applications regarding our proposal.

- *Meta-search algorithm*

In this application the real-time Internet-based technique is used as a stand-alone meta-search algorithm, capable of monitoring the continuous changes that occur in web search services, providing new and more up-to-dated results in higher ranking positions. The algorithm uses the averaged values of birth and survival rates, promoting the web search services, which adequately adapt to the web evolution and penalize those that fail to perform such information management. In other words a scoring mechanism is introduced, which its main role is to record continuously the ability of each used web search service to maintain high freshness rates due to new entries, to exclude dead links and continuously check the validity of its provided results [5].

- *Estimation of the size of the indexable web and the internet itself*

Since it is difficult to study the Internet and the web globally, we can try to evaluate his evolution through measurements from the public indexable web. The public indexable web is fixed as the part of web that is considered as indexed by most web search services. Based on the conclusions made for the portion of the indexable web covered by most of the popular search engines, the goal is to produce an estimate of the size of the entire indexable web [6]. The evaluation of the size and the evolution of the world wide web are difficult not only due to its size, but also due to its dynamic nature. It should be also taken into consideration that the web is increased exponentially in order to derive to a reliable estimate of its size. This increase is directly corresponded with the new created web pages, the web pages, which are removed, as well as the web pages that change.

For estimating the size of web in respect to the disseminated information as well as in respect to the involved network elements, many statistical methods were proposed and discussed. Among them several studies measure the size of the Internet by the number of computers connected to it, the traffic they carry as well as the number of connected nodes and their correlation. Other approaches measure the size of the disseminated data as well as the population of the web pages. By employing our proposal, the problem of measuring the active web pages of a categorized population on the world wide web at any given point in time is put into the context of capture-

recapture experiments used in wildlife biological studies as following. Web pages are to be considered as animals living on the web and the specific types of web pages, as particular species of animals whose abundance, birth and survival rates is estimated. In the web experiments, abundance corresponds to the number of active web pages, while a birth occurs every time a web page becomes active and a death occurs every time a web page becomes inactive [7].

5. Conclusions

This paper proposes a real-time Internet cache monitoring mechanism capable of monitoring the ability of the web search services to maintain new and up-to-date cache contents in their directories. The method was based on capture-recapture experiments used in real wildlife biological studies. Through a case study of sixteen-week period real-time capture-recapture experiments using the caches of two major Internet search services, we proved that our proposal is applicable for enhancing web search and measuring Internet cache evolution. The results are very interesting and give a clear idea about how differently the Internet search services update their cached content and monitor the evolution in their directories. Future works involves repeatedly evaluations and more capture-recapture sampled instances for both of the proposed applications.

References

[1] Williams, B., Nichols, J. and Conroy, M., Analysis and Management of Animal Populations, Academic Press, San Diego, California, 2002

[2] Schwarz, C. and Stobo, W., Estimating temporary migration using the robust design, Biometrics vol.53, 1997, 178–194

[3] Jolly, G., Explicit estimates from capture-recapture data with both death and immigration stochastic model, Biometrika vol. 52, 1965, 225-247

[4] Eguchi T, Statistic class notes (Topics in Ecological Statistics), Montana State University, 2000, based on book by Seber (1982; The estimation of animal abundance and related parameters. Second edition, Macmillan Publishing Co., New York, NY), downloaded from http://www.esg.montana.edu/eguchi/pdfFiles/markRecapSummary.pdf

[5] Anagnostopoulos I., Anagnostopoulos C. and Vergados D., Modeling Browsing Behavior and Sampling Web Evolution Features Through XML Instances, proceeding of the 3rd International Conference on Web Information Systems and Technologies (WEBIST 07), Barcelona, Spain, March 3-6, 2006, pp. 67-74

[6] S. Lawrence and C. L. Giles. Searching the world wide web. Science, 280(5360):98 – 100, 1998

[7] Anagnostopoulos I., Stavropoulos P., Adopting Wildlife Experiments for Web Evolution Estimations: The Role of an AI Web Page Classifier, 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 06) In: Main Conference Proceedings pp. 897-901, In: CD Proceedings: ISBN 0-7695-2750-7, Library of Congress Number 2006934379, 18-22 December 2006, Hong Kong, China