A rich internet-based programming method for measuring web freshness rates under the Pollock's sampling scheme

Ioannis Anagnostopoulos[1], George Kouzas[2], Christos – Nikolaos Anagnostopoulos[3] and Eleftherios Kayafas[2]

[1] *University of the Aegean, Department of Information and Communication Systems Engineering, 83200 Samos Island – Greece, +30 22730 82237, janag@aegean.gr*
[2] *National Technical University of Athens, School of Electrical and Computer Engineering, 15780 Zographou Campus, Athens – Greece, kayafas@cs.ntua.gr*
[3] *University of the Aegean, Department of Cultural Technology and Communications, 81100 Lesvos Island – Greece, canag@ct.aegean.gr*

***Abstract*-** In this paper, we use the Pollock's robust design method in order to measure the ability of five well known Internet search services, namely Altavista, Google, Lycos, MSN and Yahoo! (in alphabetical order), in terms of maintaining fresh and up-to-date results in their cache directories. This paper demonstrates the first Internet-based measuremets regarding the sampling of the cached results along with their initial assesment. For conducting the sampling occassions we used a client-server communication mechanism (XMLHttpRequest), which is based on the Asynchronous Javascript and XML technology (AJAX).

I. Methodology

In a capture recapture experiment the sampling process is divided into *k* primary sampling periods, each of which consists of *l* secondary sampling periods. Among primary periods the population is assumed open to gains and losses (births and deaths, emigration incidents respectively), while among secondary sampling periods the population is assumed closed [1]. During a secondary sampling period a set of species is randomly selected, marked as selected keeping a history record, and then released back to the nature. After a specific time interval the second secondary sampling periods occurs and so forth until the end of the last *l* secondary sampling period. These sampling events (secondary periods) are near and very short in time in order to assume that the populations under study are closed. This means that no losses or gains occur during these time intervals and each trapping occasions are also considered as closed. However, longer time intervals between the primary sampling periods are desirable so as births, deaths and other immigration/emigration can occur. Thus, the populations during the primary sampling periods are considered as open and thus, birth, death and survival rates are to be estimated. Figure 1 illustrates the basic structure of the Pollock's robust design method.

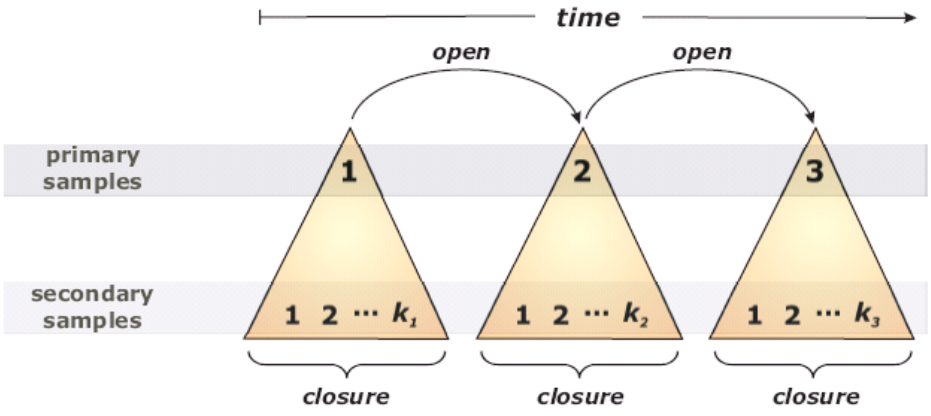

Figure 1. The basic structure of the followed capture recapture method [as appear in ref. 2]

Figure 2 depicts the proposed sampling protocol during the secondary sampling periods. The hyperlink selection is divided in two phases. In the initial phase, web pages are randomly chosen, and in the second phase, cached links are further randomly selected using the XMLHttpRequest mechanism. The two phases are described in the followings.

A. Phase 1:Random web page selection

In this phase, we randomly choose web page from the initial pool as derived from the randomly generated queries. Whenever a query is submitted, a pseudorandom number is created. The value of this number is between 0 and 1 and is compared to threshold p_1. The specific threshold is a capture – recapture variable and simulates the flipping of a coin for the web page selection. If the value is lower than threshold p_1, then the web page is selected. The web page is stored locally and is parsed in order to extract its cached results. The random number is calculated as follows:

$X = random()/Max\_Value => 0<X<1$, where X, stands for the random number

The "random()" function is a pseudorandom number generator. The range of generated numbers is between 0 and a MaxValue. All programming languages support pseudorandom generators. The stochastic ability of the system is based on system time. The task of Phase 1, is depicted in Figure 2, and simulates the coin flipping of the capture – recapture methodology.

Then, these queries are simultaneous submitted to AltaVista, Google, Lycos, MSN, and Yahoo! The first fifty results for each of the five web search services are collected, the duplicate fields are removed (keeping in parallel the search services that provided each result) and finally the merged results are stored in a local file. Then a portion of the merged-results is selected under the second phase.

B. Phase 2: Random Page Selection

Upon the completion of the first phase, the second one is initiated. The input in this phase is the results provided by the used search services in terms of cached URLs. Every time a URL is examined in this phase, a new pseudorandom number is created. The value of this number is between 0 and 1 and is compared with threshold p_2. The specific threshold is a capture – recapture variable and simulates the flipping of a coin for including the examined URL in the final pool. If the value is lower than threshold p_2, then the URL is selected and stored in a local repository. The random number is calculated as follows:

Y = *random()*/Max_Value => 0<Y<1, where Y, stands for the random number

Similarly to Phase 1, Phase 2 is depicted in Figure 2.

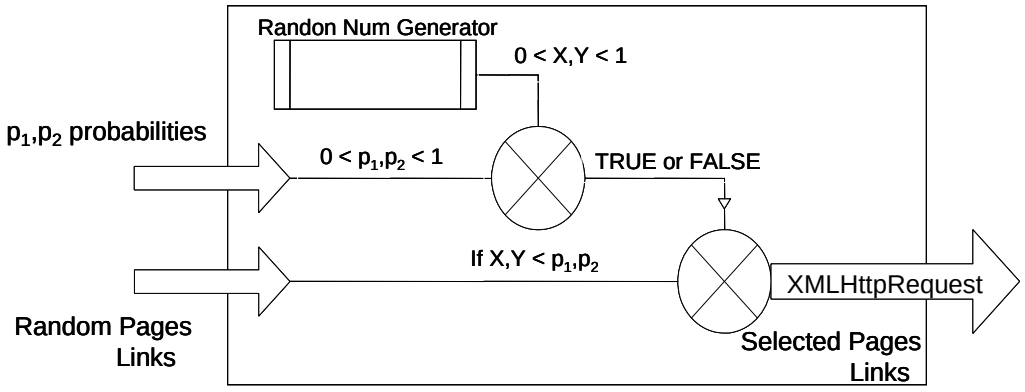

Figure 2. Random sampling generator with the AJAX-based technology

With this sequential process (Phases 1 and 2), we allow each sampling instance to have equal probability values of being included in each sample, independently of the instances that have already been sampled. This probability value is given by the product $p_1 \cdot p_2$, satisfying a fundamental assumption, which requires that selected URL in the population (either marked or unmarked) during the time of a sampling occasion *i*, must have the same probability of being captured ($p_i = p_1 \cdot p_2$, where $i = 1, 2, \ldots, k$).

Furthermore, each sampled URL (cache link) is labeled with four attributes. These are the primary and the secondary sampling period, which belongs to, the URL of the sampled result, and finally the identifier of the web search services that provide the captured result. The time intervals for the primary

are the secondary sampling periods as well as the probability values p_1 and p_2 are fine-tuned during pilot executions.

These attributes are grouped according to each of the five web search services in two main classes. The first class involves the primary indicators, which are relevant to the current primary sampling period, while the second class involves the relative indicators, which are relevant to the relation between the current primary sampling period and all the previous ones. Equations 1 and 2 express the relations between the primary statistical indicators in case where two secondary sampling periods are conducted for each primary one. Moreover, the amount of the URLs appeared in the sampling set of the current primary sampling period *i* and not appeared in any of the previous primary sampling periods are defined from Equation 1. In other words, this equation expresses the amount of the new URLs appeared in a primary sampling period *i (new{i})*, as the sum of the new URLs appeared only in the first secondary period *(new{i}$_1$)* along with the URLs that appeared only in the second secondary period *(new{i}$_2$)* and the new URLs that appeared in both of the first and second secondary period *(new{i}$_{12}$)*. Additionally, Equation 2 defines the total amount of URLs results appeared in a primary sampling period *i (total{i})* as the sum of total URLs appeared only in the first secondary period (*total{i}$_1$*) along with the total URLs that appeared only in the second secondary period (*total{i}$_2$*) and the total URLs appeared both in the first and second secondary period (*total{i}$_{12}$*).

$$new\{i\} = new\{i\}_1 + new\{i\}_2 + new\{i\}_{12} \quad (1)$$

$$total\{i\} = total\{i\}_1 + total\{i\}_2 + total\{i\}_{12} \quad (2)$$

In accordance to these two equations, if we conduct three secondary sampling occasions per one primary sampling period *i*, then the respective primary indicators for Google, are defined by Equations 3 and 4. Figure 3, illustrates a schematic representation of Equation 4 where sets A, B and C correspond to the new results of Google during the first, second and third secondary sampling periods respectively, over the total amount of sampled Google results *G_total{i}* (set T) during the primary sampling period.

$$\begin{aligned} G\_new\{i\} = {} & G\_new\{i\}_1 + G\_new\{i\}_2 + G\_new\{i\}_3 \\ & + G\_new\{i\}_{12} + G\_new\{i\}_{23} + G\_new\{i\}_{13} \\ & + G\_new\{i\}_{123} \end{aligned} \quad (3)$$

$$\begin{aligned} G\_total\{i\} = {} & G\_total\{i\}_1 + G\_total\{i\}_2 + G\_total\{i\}_3 \\ & + G\_total\{i\}_{12} + G\_total\{i\}_{23} + G\_total\{i\}_{13} \\ & + G\_total\{i\}_{123} \end{aligned} \quad (4)$$

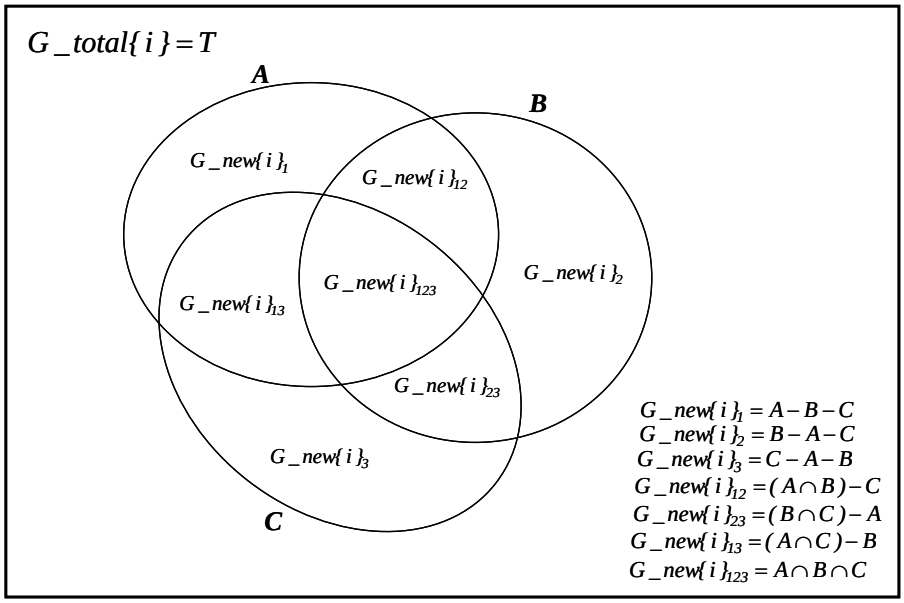

Figure 3. An example of capture recapture primary indicators and their relationships

On the other hand, during the n^{th} primary sampling period, *n*-1 set of measurements relate each current primary sampling period with its previous ones, according to the relative indicators. Equation 5 expresses the association between the relative statistical indicators in case where two secondary sampling periods are conducted for each primary one. In other words, Equation 5 defines the total amount of URLs that appeared in the current primary sampling period *n* and have also appeared in the h^{th} primary sampling period *(total_rel{nh})*. This amount is defined as the amount of the *total_rel{nh}* URLs that appeared only in the first secondary sampling occasion of the n^{th} primary sampling period $(total\_rel\{n\}_1)$ and the amount of *total_rel{nh}* URLs that appeared only in the second secondary sampling occasion of the n^{th} primary sampling period $(total\_rel\{n\}_2)$, along with the amount of the *total_rel{nh}* URLs that appeared in the sampling set of both secondary sampling occasions of the n^{th} primary sampling period ($total\_rel\{n\}_{12}$).

$$total\_rel\{nh\} = total\_rel\{n\}_1 + total\_rel\{n\}_2 + total\_rel\{n\}_{12} \quad (5)$$

Table 1. Capture recapture relative indicators (2 secondary per 4 primary sampling periods)

Current primary n	Previous primary h	Total_Rel nh	Total 1	Total 2	Total 12	search service
2	1	11	6	5	0	Altavista
2	1	5	3	2	0	Google
2	1	16	9	7	0	Lycos
2	1	7	1	6	0	MSN
2	1	7	3	4	0	Yahoo!
3	2	15	4	9	2	Altavista
3	2	12	2	7	3	Google
3	2	10	4	5	1	Lycos
3	2	9	4	3	2	MSN
3	2	11	6	5	0	Yahoo!
3	1	14	3	9	2	Altavista
3	1	9	3	5	1	Google
3	1	5	2	3	0	Lycos
3	1	13	4	6	3	MSN
3	1	7	5	2	0	Yahoo!
4	3	16	6	6	4	Altavista
4	3	11	6	5	0	Google
4	3	6	3	3	0	Lycos
4	3	12	6	4	2	MSN
4	3	6	3	2	1	Yahoo!
4	2	16	6	7	3	Altavista
4	2	14	8	4	2	Google
4	2	8	2	5	1	Lycos
4	2	14	5	7	2	MSN
4	2	11	7	4	0	Yahoo!
4	1	24	10	9	5	Altavista
4	1	9	4	4	1	Google
4	1	9	3	5	1	Lycos
4	1	5	3	2	0	MSN
4	1	9	5	4	0	Yahoo!

II. Case Study - Results

Table 1 depicts the relative indicators regarding a pilot capture recapture experiment, which consisted of four primary sampling periods and two secondary sampling periods per primary one for a specific query. Shadowed records correspond to measurements taken for Google. In these records, the third column (Total_Rel nh) indicates the total amount of identical Google results that were found in both second and first, third and second, third and first, fourth and third, fourth and second, and fourth and first primary sampling periods respectively. For example, the third column of the second shadowed record indicates the total amount of Google results that appeared in the third primary sampling period and were also last appeared in the second primary sampling period.

This amount is given according to Equation 5, as it was previously described in section II, and it is equal to $G\_total\_rel\{32\} = G\_total\_rel\{3\}_1 + G\_total\_rel\{3\}_2 + G\_total\_rel\{3\}_{12} = 2+7+3 = 12$, where $G\_total\_rel\{3\}_{12}$ is the total amount of the $G\_total\_rel\{32\}$ results that appeared in both

secondary periods of the third sampling period (column: $Total_{12}$), while $G\_total\_rel\{3\}_1$ and $G\_total\_rel\{3\}_2$ is the total amount of the $G\_total\_rel\{32\}$ results that appeared only in the first and only in the second secondary period of the third sampling period (columns: $Total_1$, $Total_2$, respectively).

Consequently, after multiple primary sampling periods and taking into account the total amount of captured instances in a current sample, the previously marked as well as the marked sampling instances in subsequent sampling periods, the proposed web evolution adaptation mechanism estimates the birth rate and the survival rate as described in [3].

III. Conclusions

The capture-recapture simulation measurements for each Internet service were further analyzed using a statistical package under the Pollock's robust design capture-recapture methodology. After the end of the experiments, the derived results have shown that during the period where the experiments were conducted the average birth rate for Google cache was higher in comparison with the other caches. This means that Google manages to provide us with more new results. In terms of birth rates Google was followed by MSN and Yahoo!, while Altavista and Lycos were far away lower.
On the other hand, MSN and Google presented the higher survival rates, presenting values, which were very close. This means that both of them have the better refresh rates and present more up-to-date results to their users. Once again Yahoo! followed at the third place (not far away), while Altavista and Lycos presented lower average survival rates. The use of the AJAX-based web programming technology and the XMLHttpRequest mechanism, helped us to acquire faster and more reliable samplings.

IV. Acknowledgements

The authors would like to thank Dr. Photis Stavropoulos for his valuable help and comments, in related research done during last years.

References

[1] Pollock K.H., Nichols J.D., Brownie C., Hines J.E., Statistical inference for capture-recapture experiments, Wildlife Monographs 107, 1990.
[2] Program MARK: A gentle introduction, http://www.phidot.org/software/mark/docs/book/.
[3] I. Anagnostopoulos, P. Stavropoulos, G. Kouzas, C. Anagnostopoulos, D.D.Vergados, Estimating the evolution of categorized web page populations, In Proc. of the 1st International Workshop on Adaptation and Evolution in Web Systems Engineering (AEWSE 06), 6th International Conference of Web Engineering (ICWE 06), Palo Alto, CA, July 10-14, 2006, ISBN: 1-59593-435-9.