Monitoring the Evolution of Cached Content in Google and MSN

Ioannis Anagnostopoulos
University of the Aegean, Department of Information and Communications Systems Engineering
Karlovassi - 83200, Samos, Greece
+30 22730 82237
janag@aegean.gr

ABSTRACT
In this paper, we describe a capture-recapture experiment conducted on Google's and MSN's cached directories. The anticipated outcome of this work was to monitor evolution rates in these web search services as well as measure their ability to index and maintain fresh and up-to-date results in their cached directories.

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *web-based services,* H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

General Terms: Algorithms, Measurement, Performance

Keywords: capture-recapture methodology, web cached content, internet evolution rates

1. WEB CAPTURE-RECAPTURE EXPERIMENTS
Our experiments are put into the context of capture-recapture experiments used in wildlife biological studies. 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 the appropriate models one can estimate the total population size, as well as the birth rate, the death rate and the survival rate of each species. The sampling scheme chosen for capturing, marking and recapturing the cached web pages is the robust design, which extends the Jolly-Seber Model [1]. 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 the web paradigm, death corresponds to a result that is no longer exists (dead links, errors 404), birth match to a new result (new or updated information), while incidents of immigration and/or emigration correspond to active but temporary unavailable results (e.g. errors of type 50*, web server internal errors, bad gateway, service/host unavailable, etc). The necessary amendments and modifications made in order to conduct capture-recapture measurements based on the real-life experiments are described in [2].

During September of 2006 and January of 2007 we conducted a sixteen-week period capture-recapture experiment using the indexes of Google and MSN. For acquiring estimations using capture-recapture measurements in nature, the researcher needs at least two primary sampling periods, which each one consist of at least two secondary sampling periods. In our case, we divided our experiments in eight primary sampling periods, which each one consisted of eight secondary sampling periods. In order to calculate the time-interval between two primary sampling periods we conducted a pretest, which lasted for ten days. During this period we checked the refresh rate of Google and MSN. The refresh rate was calculated by averaging the differences between the days of the experiments and the dates where the web search engines' had last updated their indexes for the first ten results of Google and MSN.

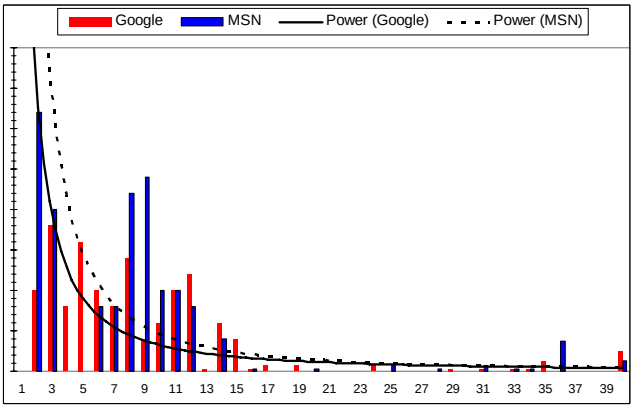

Figure 1. Daily up-to-date rates (pretest - relative values)

Thus, during the pretest we submitted 200 random three-term queries in Google and MSN using the Mangle Random Link Generator (available at http://www.mangle.ca/randomweb/). This means that we processed nearly 4000 results (200 queries x 2 search engines x 10 top results). After the pretest we noticed that almost the half amount of the results that Google provides are refreshed during a week-time period or less. For the same period MSN refresh at least the 43% of its results. However, it is worth to notice that during the pretest the portion of the refreshed results of MSN that were refreshed within a time-span of less than three days was 32% in respect to nearly 20% of Google. The respective amount of returned results that were refreshed during two, three, and four weeks or less for Google-MSN were 88.4%-92.4%, 94.2%-92.8%, and 95.6%-93.7% respectively. Figure 1 illustrates the distribution of the up-to-dateness for all the examined results on a daily-scale for Google and MSN, as well as their power trendlines. Having examined all top-ten results for the 200 randomly submitted queries, it was calculated that these results were being updated by Google and MSN within an average frequency of 9.11 and 8.79 days respectively.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

Taking into consideration the above measurements and since more than 90% of the retuned results from both Google and MSN were refreshed within a time-span of less than two weeks, the time interval between two subsequent primary sampling periods was chosen to be 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. Having created a pool of 100 randomly formed tree-term queries in English (different form the pretest set) and as Figure 2 illustrates, during each secondary sampling period a portion of these queries was submitted according to a probability values p_1 (step 1). In our case p_1 was set equal to 0.3. After the completion of this procedure for all N (=100) queries, the selected amount of queries (approx. p_1*N) were submitted in Google and MSN (step 2). Finally in step 3, we checked two main issues. At first we examined the differences between the days of the experiments and the dates where the web search engines' had last updated their indexes for the examined top-fifty results of Google and MSN (Refreshness). In parallel, we check the ability of Google and MSN in terms of maintaining in their caches, the up-to-date content, which is disseminated on the web, for all the top-fifty tested results (Up-to-dateness). In our case, since we decided to include all the top-fifty returned results per search service used (p_2 and T equal to 1 and 50 respectively). The procedure is repeated from step 3 to step 1, until the final secondary sampling period is completed.

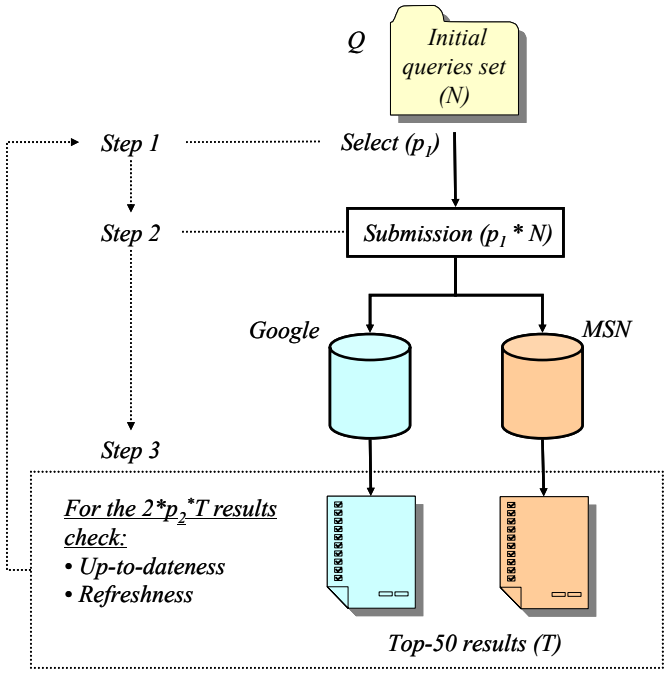

Figure 2. Steps during subsequent secondary sampling periods

2. RESULTS - CONCLUSIONS

Table 1, holds all the respective parameters regarding the capture-recapture measurements, which was conducted from September of 2006 to January of 2007 (16 weeks). Thus, *Ni* stands for the absolute values of the tested results for Google and MSN, 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 Google included 485 new results in its index (B1) over 23396 total provided results (N1+N2), while MSN 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 Google and MSN 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% (Google and MSN).

Finally, for the whole period of the experiments (16 weeks), Google presented (in average values) larger levels in birth rates and lower levels in survival rates. This means that Google not only indexed more new results ($avg(b)_{Google} = 0.0402 > avg(b)_{MSN} = 0.0347$), but also manage to maintain a larger amount of results where their actual disseminated content was the same with the cached content during the experiments ($avg(\varphi)_{Google} = 0.9603 < avg(\varphi)_{MSN} = 0.9652$). 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, even if that, between subsequent sampling periods one was sometimes present better rates over the other.

Table 1. Capture-recapture measurements

N: Active population (absolute) B: Births (absolute), b: Birth rate φ: Survival rate					
	Google	MSN		Google	MSN
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

3. REFERENCES

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

[2] Anagnostopoulos, I. and Stavropoulos, P. Adopting Wildlife Experiments for Web Evolution Estimations: The Role of an AI Web Page Classifier, In Proceedings of 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 06), pp. 897-901, 18-22 December 2006, Hong Kong, China.