XVII IMEKO World Congress
Metrology in the 3rd Millennium
June 22–27, 2003, Dubrovnik, Croatia

MEASURING THE POPULATION OF WEB PAGES IN THE WILD WEB

Georgios S. Kouzas, Photis Stavropoulos, Ioannis Anagnostopoulos, Christos Anagnostopoulos, Vassili Loumos and Eleftherios Kayafas

National Technical University of Athens
Department of Electrical and Computer Engineer
Heroon Politechneiou 9 Str, Zographou Athens - Greece

Abstract − This paper describes a system capable of measuring the type of pages in the wild WEB. The system uses as theoretical background the capture – recapture methodology, which is mainly used for statistical purposes in other scientific areas. The problem addressed may be compared to the context of capture – recapture experiments used in wildlife biological studies. The novelty in the proposed method stands in the fact that the experiment takes place in the wild WEB and instead of animal species, the measurements concern several types of WEB pages.

Keywords: statistical measurements, WEB pages, capture – recapture methodology.

1. INTRODUCTION

A rapid growth of Internet activity is observed recently, especially concerning WEB applications. A similar development in Internet technologies follows this revolution. When the first WEB site was launched, the content was disseminated only in static HTML format. Since then, new technologies were invented, matured and died. Nowadays, when a new company intends to enter the WEB, it probably needs to know which is the appropriate technology (based on subjective criteria) to be used for the creation of its WEB site. One of the most important criteria is the usage of each technology at the current period. However, due to the vast expansion of the WEB, it is impossible to measure the exact number of WEB pages created using each technology. The present methodology proposes a system with the ability to calculate the percentage of usage for each WEB technology.

2. METHODOLOGY

The basic idea is to measure the estimation of population for each page type through the capture-recapture methodology, instead of measuring Internet technologies. This means that WEB pages are regarded as animals living in the wild and a specific WEB page technology is regarded as a particular animal species, whose abundance, birth and survival rates should be estimated. Abundance is the number of pages created by a specific technology; a birth occurs every time a new page becomes active and a death occurs every time a page becomes inactive. The idea behind the capture-recapture experiments is to set up some traps in an area where the species under study lives, catch some individuals, mark them and then release them. Some time later, the traps are set up again and a number of animals are caught. By specifying how many individuals were previously marked, the number of existing animals in this species, can be estimated. The same is applied to WEB pages.

The sampling scheme chosen for this purpose is referred in [1]. Every week, the same time, a primary sampling period commences, during which, two possibly overlapping samples of WEB pages are randomly selected. The moments at which the two samples are drawn are the secondary sampling periods. These two secondary periods are so close that no new pages become active or inactive between the first and the second sample selection. The data derived from the two secondary periods of each primary one, permit accurate and precise estimation of the number of active WEB pages (of each technology) at that point in time.

The primary sampling periods**,** on the other hand, are set in a week order of magnitude, because this is enough time for the population of pages to change to some extent. The data from each pair of consecutive primary periods allow, besides the estimation of the population size, the estimation of the birth and survival rates for each type of WEB page technology. The process can be sustained for as many primary-sampling periods as desired.

In each sample, the pages for each technology are identified. The best way to identify each technology is by examining the extension of the tested URL. For example, the page http://www.domain.com/page.asp was created using Active Server Pages (ASP). Among the pages of each technology the number of those that have been previously marked and of those, which are encountered for the first time is estimated. The latter is also marked to facilitate their identification in future samplings. The data consists of the numbers of different WEB page technologies and is then analyzed using the statistical methodology [2]. There are two steps in order to recognize the population estimation for each technology in the proposed system:

- the sampling phase
- the statistical population estimation for each WEB page technology

1. PROPOSED SYSTEM

Figure 1 depicts the proposed system, which consists of two main parts. The first part is responsible for data collection (the sampling phase) and the second one for calculating the statistical indicators (the statistical phase). According to the capture – recapture methodology, the birth and death rates, in other words the final population estimation, is extracted from these indicators.

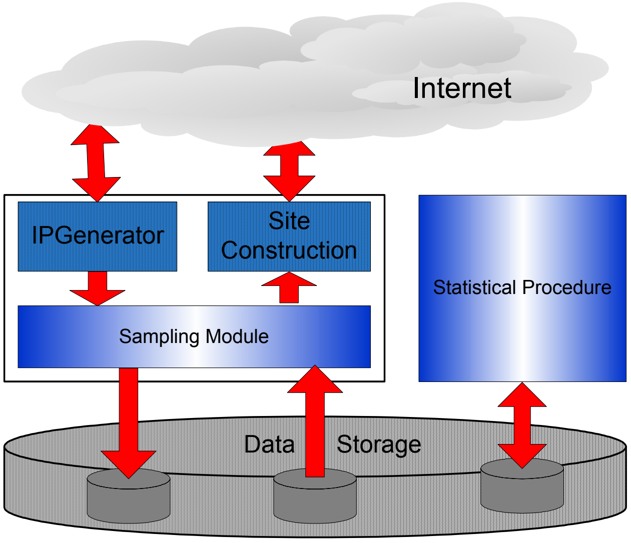


Figure 1: Proposed system

3.1. Sampling Phase

During the sampling phase, the system collects pages from the Internet and stores them locally. Based on the capture – recapture methodology, the page collection must be random. Thus, firstly a module generates and validates random IP addresses. Every valid IP address corresponds to an active WEB site. Next, a number of valid IP addresses, is chosen randomly with a probability value p1 and their site trees are extracted. Finally, some WEB pages are selected randomly from each site tree, with a probability value p2 and are stored into a local database.

The three main procedures of the sampling phase are:

- The IP Generator:
- The Site construction
- The Sampling module

3.1.1. The IP Generator

This procedure is used as a peripheral module to the whole system. It has the responsibility of producing a set of valid IP numbers that correspond to usable domain names. The task here is to collect an amount of domain names and insert them on a local Database for future use.

The generated IP numbers must be produced randomly. IP numbers consist of 4 (8 bit) bytes -giving a total of 32 bits of available information. The IP number *147.102.16.32* is an example of this: 4 (decimal) numbers separated by (.) dots. The IP generator produces four random numbers to fill the four different fields. So the produced IP number has the following format:

*Rand_num1***.***Rand_num2***.***Rand_num3***.** *Rand_num4*

Each of the four numbers has the same probability to be produced, so that favored IP addresses do not exist. The random numbers produced are between 1 and 254, so the possible set of IP productions is {1.1.1.1 -> 254.254.254.254}. With a simple calculation, this set has 254^4 possible IP numbers that are about 4 billion numbers. From this set, some special IP addresses are excluded because they are reserved for 'unconnected' methods. For instance, the IP 127.0.0.0 is a network mask and not an IP number - it is used to modify how local IP numbers are interpreted locally so there is no point in producing this kind of IPs. The validation phase of the IP Generator is a typical http request to the produced IP. If this request responds with an error, it means that the IP is invalid and that a WEB server does not exist behind it. Otherwise, the produced IP is stored. An approximation for the number of valid IPs, is about 2 valid per 1000 IPs produced. So, as shown in Figure 2, the multithread programming technique was used for the validation phase, in order to reduce the total time of validation.

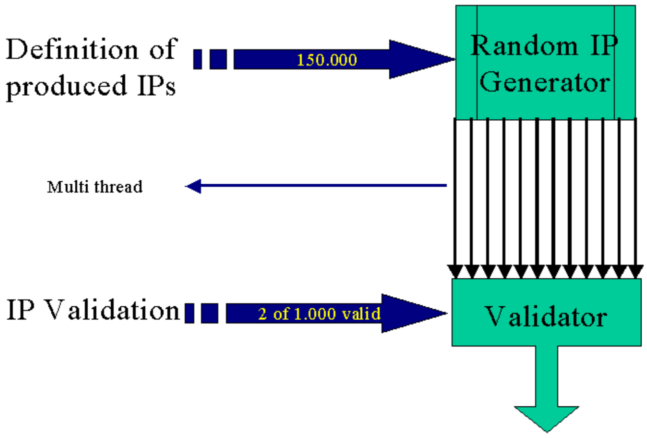


Figure2: IP Validator

3.1.2. The Site Construction

The scope of the tree extractor is to construct a tree representation consisting of the internal links that a specific domain may include.

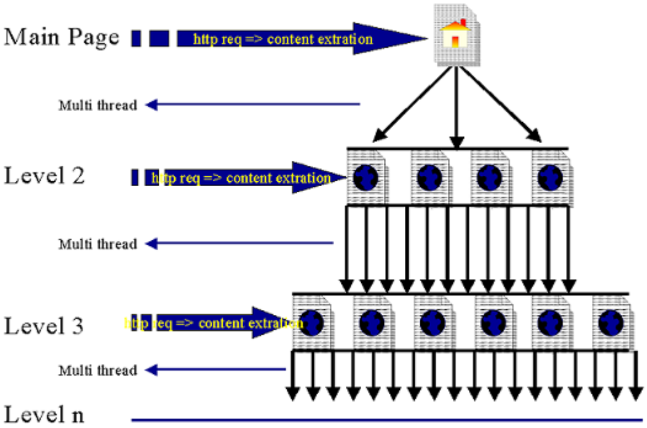

Main Page
http req => content extration
Multi thread
Level 2
http req => content extration
Multi thread
Level 3
http req => content extration
Multi thread
Level n

Figure 3: Site Construction Algorithm

The main idea is to extract the html content of a specific WEB page and to retrieve the child pages under the initial page. A multilevel extraction technique, depicted in Figure 3, was used for the tree extractor.

In the first level, the page URL, the page type and the content of the home page of the site are stored. After the content parsing, all child URLs are extracted. It must be noticed that external links are ignored. All extracted URLs from the first level are the input for the second one. So, in the second level, the same procedure is executed, for each new WEB page. To reduce the time of the extraction (for large sized WEB Sites, the extraction time could reach one day in order of magnitude), the multithread technique was applied to every level. This technique is repeated for "n" levels until the whole tree is extracted.

3.1.3. The sampling module

The WEB page sample selection is divided in two phases. In the initial phase, sites are chosen randomly, and in the second phase, WEB pages, which belong to the selected sites, are also randomly selected. The WEB technologies, which are going to be measured, are stored in an external file as extensions. Thus, for example if the proposed system is going to measure two technologies (ASP and HTML), the two corresponding extensions will be put on the "extensions.dat" file. The sampling module reads this file and categorizes the pages to the appropriate WEB technology.

Phase 1:Random site selection

In this task, the system is fed with the IPs or domain names, which correspond to a site. The IP Generator module is responsible for feeding the random site selection phase, as described previously. Whenever an IP Generator initiates the task, a pseudorandom number is created. The value of this number is between 0 and 1 and is compared to threshold p1. The specific threshold is a capture – recapture variable and simulates the flipping of a coin for the site selection. If the value is lower than threshold p1, then the site is selected. The site URL is stored locally and is parsed to the treeExtractor module, which extracts the site tree. The random number is calculated as follows:

X = random()/Max_Value => 0<X<1

Where X, 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.

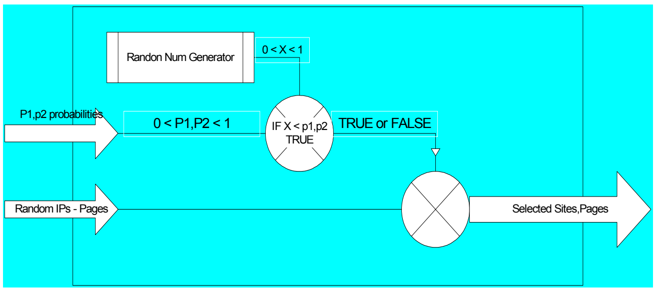


Figure 4: Algorithm for Site/Page selection

The task of phase 1, depicted in Figure 4, simulates the coin flipping of the capture – recapture methodology.

Phase 2: Random Page Selection

When the first phase is finished, the second one, which selects pages randomly, commences. The input for this task is the tree of a selected site, fed with pages from the Site Construction procedure. Every time the Site Construction procedure feeds this task with a WEB page, a new pseudorandom number is created. The value of this number is between 0 and 1 and is compared with threshold p2. The specific threshold is a capture – recapture variable and simulates the flipping of a coin for WEB page selection, similar to threshold p1. If the value is lower than threshold p2, then the page is selected. The WEB page URL is stored in a local database.

The random number is calculated as follows:

Y = random()/Max_Value => 0<X<1

Similarly to the task of phase 1, the task of phase 2 is depicted in Figure 4.

3.2. Statistical phase

This procedure creates the specific indicators for the capture - recapture methodology. After statistical processing, these indicators extract the final estimation of WEB page technologies. The specific data is retrieved from the local database.

The general input type for each data row has the following format:

- Primary sampling period
- Secondary sampling period
- Selected Page
- Site of the selected page

The statistical indicators produced with this module are grouped into two main classes:

- The first class (**primary indicators class**) includes indicators, relevant to the current primary sampling period.
- The second class (**relative indicators class**) includes indicators, which refer to the relation between the current primary sampling period and all the previous ones.

3.2.1. Primary indicators class

In every primary sampling period, a row (set) of these statistical results is created for every WEB page technology. Every row contains the following fields:

- NT: The number of pages that appeared in the sampling set of the last sampling period. These pages should not have appeared in any of the previous primary sampling periods.
- NB: The number of new pages that appeared in the sampling set of both the first and the second secondary sampling period.
- NF: The number of new pages that appeared in the sampling set of the first secondary sampling period but did not appear in the sampling set of the second secondary sampling period.
- NS: The number of new pages that appeared in the sampling set of the second secondary sampling period but did not appear in the sampling set of the first secondary sampling period.

- PT: The number of pages that appeared in the last primary sampling period and also appeared in at least one previous primary sampling period. In other words, old pages, which appeared in the last primary sampling period.
- TT: The total number of pages, which appeared in the sampling set of the last primary sampling period.
- TB: The total number of pages, which appeared in the sampling set of both the first and the second secondary sampling period of the last primary sampling period.
- TF: The total number of pages that appeared in the sampling set of the first secondary sampling period but did not appear in the sampling set of the second secondary sampling period.
- TS: The total number of pages that appeared in the sampling set of the second secondary sampling period but did not appear in the sampling set of the first secondary sampling period.

The relations between the aforementioned statistical results can be expressed with the following equations:

$NT = NB + NF + NS$

$TT = TB + TF + TS$

$TT = PT + NT$

3.2.2. Relative indicators class

In every primary sampling period, a set of rows with these statistical results is created for every WEB page technology. If the number of the primary sampling period is “n” (where n is a natural number), n-1 rows with statistical results are created. Every row contains the following fields:

- H_T: The number of pages that appeared in the sampling set of the last primary sampling period and last appeared in the h primary sampling period.
- H_B: The number of HT pages that appeared in the sampling set of both first and second secondary sampling periods of the last primary sampling period.
- H_F: The number of HT pages that appeared in the sampling set of the first secondary sampling period but did not appear in the sampling set of the second secondary sampling period of the last primary sampling period.
- H_S: The number of HT pages that appeared in the sampling set of the second secondary sampling period but did not appear in the sampling set of the first secondary sampling period of the last primary sampling period.

From the above it is calculated that:

$HT = HB+HF+HS$

4. MEASUREMENTS

Before the initiation of the system, the following variables should be fine-tuned.

- Probability p1
- Probability p2
- Time duration between two secondary periods (T1)
- Time duration between two primary periods (T2)
- Number of produced random IPs.

In order to produce reliable results, the two secondary sampling periods should have an overlapping percentage. This requires that the sample set of sites and pages in each (secondary) sampling period are quite large. Thus, the total number of produced IPs should be defined to a point that the system is able to produce a satisfactory set of selected sites and pages. For the same reason, p1 and p2 must be defined to an appropriate value.

Another crucial point is that the time duration between the two secondary periods must be short enough, so that the population of each type is considered invariable. For this purpose, the aforementioned time duration was set to zero.

Finally, the time duration between two primary sampling periods should be large enough, so that changes can occur in the population. The initial value was set to 10 days. For the final definition of each variable, the proposed system must be tested a number of times, in each machine it is installed. The parameter that affects these variables is the network speed of the target machine.

The proposed system was tested for several sampling periods. The values of the five aforementioned variables, were: 150.000 randomly produced IPs, p1=p2=0,8, T1=0 T2=10days, and the network connection was 1Gbps. The tested WEB technologies were five (ASP, JSP, PHP, PL and HTM*).

The indicators shown in Table 1 and Table 2, refer to the first four primary sampling periods. Each primary period was tested for 150.000 random IPs and about 300.000 pages were extracted. The WEB type technology of each selected page was recognized and stored in a local database. With the aid of a statistical program, one can extract from these indicators the final estimation population, as well as the birth and death rates for each WEB technology.

In Table 1, the primary indicator class is shown. The first column refers to the primary sampling period, the last column to WEB technology and the other columns refer to the indicators explained above.

In Table 2, the relative indicator class is shown. Again, the first column refers to the primary sampling period, the last column to WEB technology and the other columns refer to the indicators explained above.

5. CONCLUSIONS

This paper describes a system capable of measuring WEB pages in an automatic way from the wild WEB. A real experiment is performed based in the well-known methodology of capture – recapture and its produced indicators. The anticipated outcome of this system is to estimate the type of WEB pages that disseminate their content in the Internet.

After examining a large training set, some limitations got out, that need further research. Although multi-thread technology was used, the tree extraction of each WEB site took a long time to be completed. This increased the total time for each secondary sampling period (one day in order of magnitude), which is opposed to the capture – recapture methodology that requires short time duration for each secondary period. The problem concentrates to the network bandwidth, which is incapable of accepting a large number of http requests. Another point is that more than one WEB site may exist behind of each IP address. Considering each IP address as a different WEB site, some WEB sites may be lost. Using the dnslookup() function, to reserve IP addresses to the corresponding domain names, failed because many sites did not have the reserve mirror of DNS.

For future work, the proposed system could be combined with a module capable of classifying special types of WEB pages. For example, a combination with a neural network, such as the proposed system in [4], which is capable of classifying e-commerce pages, will have the ability to calculate an estimation population of e-commerce pages on the WEB. An implementation of a system like the above was tested and evaluated for the scopes of the European founded Project (ERMIS "Electronic commeRce Measurements through Intelligent agentS – IST - 1999-21051) [3].

6. AKNOWLEDGEMENTS

This system was tested and evaluated for the scopes of the European founded Project (ERMIS "Electronic commeRce Measurements through Intelligent agentS – IST - 1999-21051) [3].

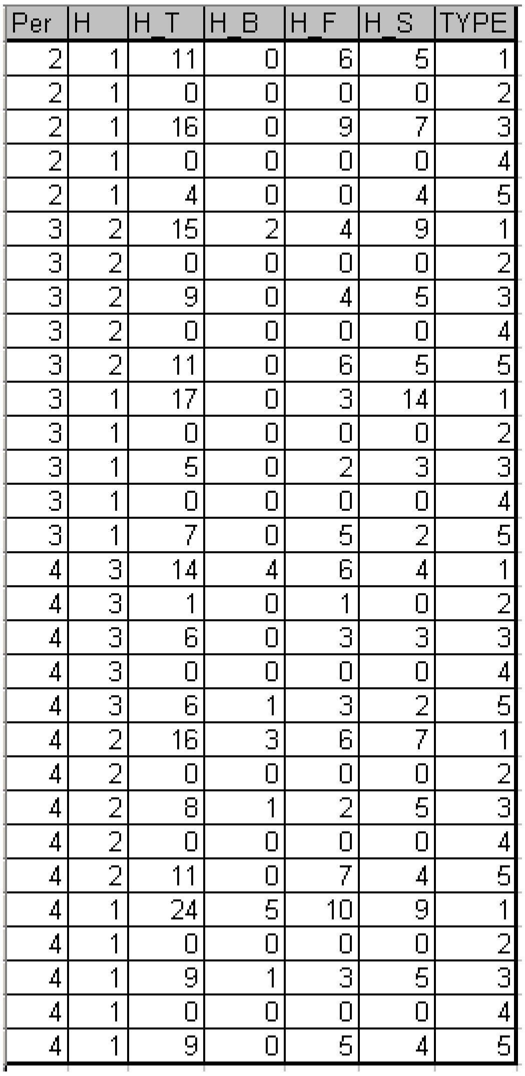
REFERENCES

[1] Kendall, W. L., Pollock, K. H. and Brownie, C. (1995). A likelihood-based approach to capture-recapture estimation of demographic parameters under the robust design. *Biometrics*, 51, 293-308.

[2] Draper, D. and Bowater, R. Sampling errors under non-probability sampling. Davies, P. and Smith, P. (eds) *Model quality report in business statistics*, Eurostat.

[3] Statistical Methodology for measurement, WP1, Deliverable D1.3, ERMIS Consortium, http://www.ermisproject.gr, May 2002.

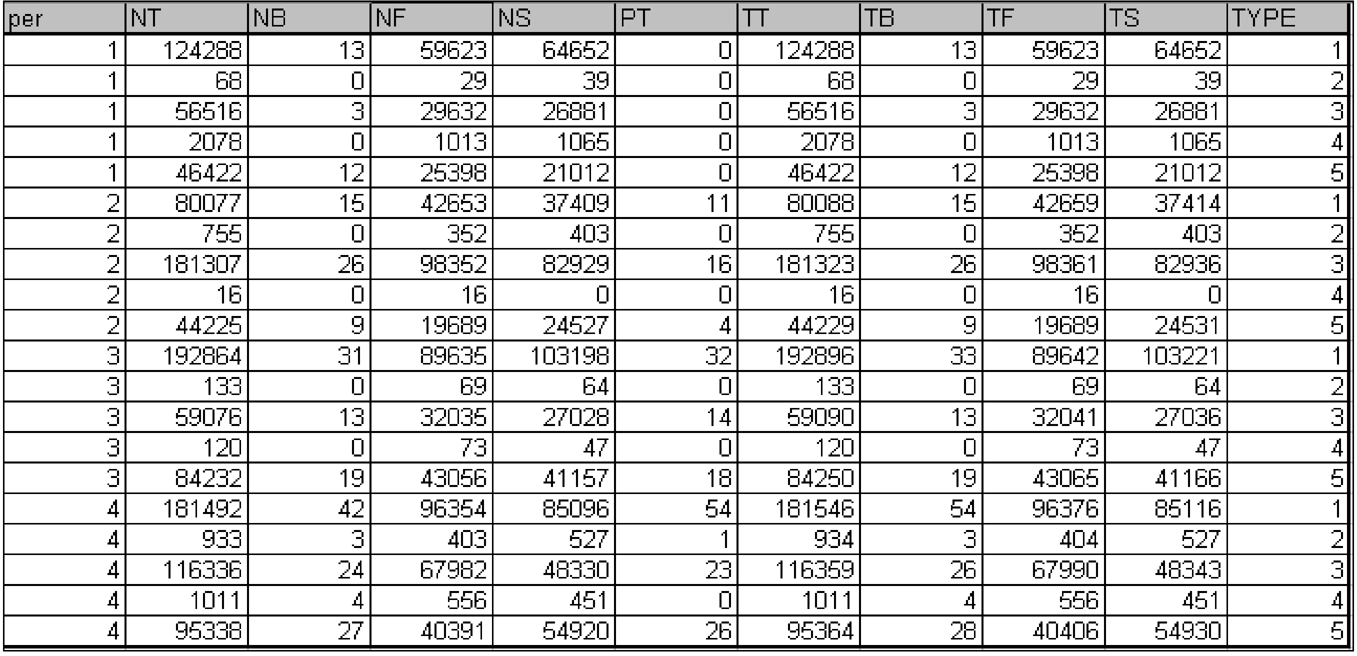
Table 1: relative indicator class



Per	H	H_T	H_B	H_F	H_S	TYPE
2	1	11	0	6	5	1
2	1	0	0	0	0	2
2	1	16	0	9	7	3
2	1	0	0	0	0	4
2	1	4	0	0	4	5
3	2	15	2	4	9	1
3	2	0	0	0	0	2
3	2	9	0	4	5	3
3	2	0	0	0	0	4
3	2	11	0	6	5	5
3	1	17	0	3	14	1
3	1	0	0	0	0	2
3	1	5	0	2	3	3
3	1	0	0	0	0	4
3	1	7	0	5	2	5
4	3	14	4	6	4	1
4	3	1	0	1	0	2
4	3	6	0	3	3	3
4	3	0	0	0	0	4
4	3	6	1	3	2	5
4	2	16	3	6	7	1
4	2	0	0	0	0	2
4	2	8	1	2	5	3
4	2	0	0	0	0	4
4	2	11	0	7	4	5
4	1	24	5	10	9	1
4	1	0	0	0	0	2
4	1	9	1	3	5	3
4	1	0	0	0	0	4
4	1	9	0	5	4	5

Authors: Mr. Georgios S. Kouzas, Mr. Photis Stavropoulos Mr. Ioannis Anagnostopoulos, Dr. Christos Anagnostopoulos, Prof. Loumos Vassili, Prof. Kayafas Eleftherios, Department of Electrical and Computer Engineering, National Technical University of Athens (NTUA), 9, Heroon Polytechneiou Str., Zographou, Athens, Greece, Tel: +30 210 772 2538, Fax: +30 210 772 2538
Contact author e-mail address: gkouzas@ece.ntua.gr

Table 2: Primary indicator class



per	NT	NB	NF	NS	PT	TT	TB	TF	TS	TYPE
1	124288	13	59623	64652	0	124288	13	59623	64652	1
1	68	0	29	39	0	68	0	29	39	2
1	56516	3	29632	26881	0	56516	3	29632	26881	3
1	2078	0	1013	1065	0	2078	0	1013	1065	4
1	46422	12	25398	21012	0	46422	12	25398	21012	5
2	80077	15	42653	37409	11	80088	15	42659	37414	1
2	755	0	352	403	0	755	0	352	403	2
2	181307	26	98352	82929	16	181323	26	98361	82936	3
2	16	0	16	0	0	16	0	16	0	4
2	44225	9	19689	24527	4	44229	9	19689	24531	5
3	192864	31	89635	103198	32	192896	33	89642	103221	1
3	133	0	69	64	0	133	0	69	64	2
3	59076	13	32035	27028	14	59090	13	32041	27036	3
3	120	0	73	47	0	120	0	73	47	4
3	84232	19	43056	41157	18	84250	19	43065	41166	5
4	181492	42	96354	85096	54	181546	54	96376	85116	1
4	933	3	403	527	1	934	3	404	527	2
4	116336	24	67982	48330	23	116359	26	67990	48343	3
4	1011	4	556	451	0	1011	4	556	451	4
4	95338	27	40391	54920	26	95364	28	40406	54930	5