ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

A novel approach to analyze and estimate error rate for a web Crawler Data Set

Rajesh. L[1*], Shanthi. V[2], Lakshmi Narasimhan.V[3]

[1] Department of CSA, SCSVMV University, Kanchipuram, India

[2] Department of MCA, St. Joseph's College of Engineering, Chennai, India

[3] Department of CSE, RGMCET, Andhra Pradesh, India

***Corresponding author: E-Mail: prof.lrajesh@gmail.com, drvshanthi@yahoo.co.in**

ABSTRACT

Web Crawlers (also called Web Spiders or Robots), are programs used to download documents from the internet. Search Engines utilizes the service of web crawler to index the web page. Web crawlers, which download specific topic oriented URLs are called Focused Crawler. Web Crawlers simply downloads all the URLs which come in their way during the crawling process. Focused Web crawlers retrieve the pages, process them, compare their data to filter out pages that are relevant to the topic of discussion and classify them before sending them to the search engine for indexing. This paper addresses the problem of estimating the error / failure rate of the focused web crawler which skips the relevant topic oriented URL from the classification process. Evaluating results using standard measures like precision and recall is difficult due to unavailable information of total number of relevant / existing documents. Authors of this paper has introduced a new metric called Mean Sample Size Weighted Failure Rate (MSSWFR) which will provide the comprehensive results to estimate the error occurred during the classification process by using several standard statistical measures. The veracity and robustness of the statistical methods employed have been verified by computing the value of MSSWFR equals 19.3402 which is close to the non-weighted mean failure rate of 18.3166. These values verify the statistical significance of the methods employed in this paper.

KEY WORDS: Kanchi Crawler, Content Metrics, Base URLs.

1. INTRODUCTION

Search Engines becomes inevitable ingredient to gather information available in the world of Internet (Zhou, 2010). For better quality search engines, the most crucial factor is the relevancy of results produced by them. A Web Crawler is an important component of any search engine. A Web Crawler recursively traverse and download web pages that can be analyzed and mined in a central location, either online (available on the web) or off-line (local storage) (Lilliefors, 1967). Web crawlers are of different types based on the nature of collecting the information. Focused crawler is one such type of web crawler which is designed in such a way that it gathers document on a specific topic (Chakrabarti, 1999). To index a document URL, the Focused crawler should ensure that the document which is under the review belongs to the specific topic.

Sampling Errors – A Literature View

Sampling Errors: The error arising due to drawing inferences about the population on the basis of few observations is termed as sampling error. Even if atmost care has been taken in selecting a sample, the results derived from a sample study may not be exactly equal to the actual value in the population. The reason is that the sampled estimate is based on the part of the population and not on the whole. Sampling errors are of two types:

Bias error: Bias error arises from any bias in the sample selection process. For example, if in the place of a systematic sampling procedure, Judgement sampling has been used then a degree of bias is introduced and hence the errors are called bias errors. Bias may be caused due to following factors:

Faulty process of Selection: Faulty selection of the sample gives rise to bias in a number of was such as Deliberate (purposively) selection of a representative sample, Substitution, Non-response and an appeal to the vanity.

Faulty collection of data: Any consistent error measurement will give rise to bias whether the measurements are carried out on a sample or on all the units of population.

Bias in Analysis: In addition to the bias which arises from faulty process of selection and faulty collection of information, faulty method of analysis also introduces bias. Such bias can be overcome by choosing the proper method of analysis.

Unbiased Error: Unbiased error arises due to chance differences between the members of the population included in the sample and those not included.

The list of errors which occurs during the crawling process in the web environment due to several reasons are listed in the Table 1

Table.1.List of errors occurred during the Web Crawling Process

Error Code	Description	Error Code	Description
204	No Content	405	Method Not Allowed
205	Reset Content	406	Not Acceptable
300	Multiple Choices	407	Proxy Authentication Required
301	Moved Permanently	408	Request Time-Out
302	Moved Temporarily	500	Server Error
303	See Other	501	Not Implemented
400	Bad Request	502	Bad Gateway
401	Unauthorized	503	Out of Resources
403	Forbidden	504	Gateway Time-Out
404	Not Found	505	HTTP Version not supported

Non-Sampling Error: The error mainly arising at the stage of ascertainment and processing of data are known as non-sampling errors, which are both common in complete enumeration and sample surveys. In the processing of data, tabulation errors may be committed affecting the final results. Thus in the census enumeration method, the data obtained although free from sampling errors, would still be subject to non-sampling errors.

2. METHODS & MATERIALS

Proposed Approach

Experimental Setup: Kanchi Crawler, a focused Web Crawler is implemented to collect the data which forms basis for the dataset. Here, the data is nothing but the collection of URLs. When the corpus size is large, it is difficult to perform detailed analysis; instead we can instrument the corpus and create a control corpus / sample (of smaller size) and perform analysis on the control corpus. It is then possible to extrapolate the analysis to the general larger corpus. The total number of URLs identified correctly or missed is calculated for each sample corpora ranging from 25k to 500k with the scaling difference of 25k with each step. Note that, we have the total corpus size of 500,000. Initially the Kanchi Crawler performed classification process with the test corpus size of 25,000 URLs. As a next step, we increase the test corpus size to 50,000 and instrumented our 500 URLs to test whether the increase in corpus scaling will have any significance for the crawler efficiency towards classifying the URLs. The corpus size is increased by adding 25000 URLs to existing test corpus and in such increments until all the URLs are accommodated by having the corpus size of 500,000 URLs. We have performed experiments on the control corpus at least 10 times for each given control corpus size for each case of control corpus and the corresponding result are shown in Table 2. KanchiCrawler missed to classify the relevant URLs which we call it as a failure rate for the corresponding corpus size. For each sample size, the Sample Means, Standard Deviations and variances related to the failure rate are computed and listed in the Table 2.

Table.2.Mean, Standard Deviations & Variances for various corpus size

Control Corpus	Sample Mean	Standard Deviation	Variance
25K	14.55	6.08	36.98
50K	15.66	5.78	33.43
75K	15.22	10.72	115.12
100K	14.88	7.73	59.82
125K	16.77	5.98	35.87
150K	14.00	5.94	35.37
175K	18.33	7.42	55.16
200K	19.22	3.88	15.12
225K	20.11	6.13	37.65
250K	21.55	8.42	70.93
275K	18.66	6.23	38.84
300K	19.11	3.85	14.84
325K	17.78	2.78	7.77
350K	18.44	3.80	14.48
375K	20.77	5.45	29.78
400K	19.44	5.21	27.16
425K	18.22	5.58	31.15
450K	20.33	6.37	40.67
475K	23.22	5.30	28.10
500K	20.00	5.14	26.44

Let us consider Table 2 which details values of control corpus size range from 25k to 500k. The table lists the values of corpus size, sample mean, standard deviation and variance. For instance on row 6, for a corpus size of 150k the sample mean is 14, with the standard deviation of 5.94 and variance of 35.37.

Table.3.Normalized Mean Table

Test Corpus	Optimal Normalized Mean	Normalized Mean
25K	18.60	16.9267
50K	18.60	18.3175
75K	18.60	17.8063
100K	18.60	17.3823
125K	18.60	19.2874
150K	18.60	16.1053
175K	18.60	19.8003
200K	18.60	19.6212
225K	18.60	19.1061
250K	18.60	17.6267
275K	18.60	19.7735
300K	18.60	19.6623
325K	18.60	19.7367
350K	18.60	19.7968
375K	18.60	18.5146
400K	18.60	19.5232
425K	18.60	19.7984
450K	18.60	18.9276
475K	18.60	15.1643
500K	18.60	19.1881

Let us consider the Table 3 which details normalized mean deficiency. The table lists the values of size of control corpus, optimum normalized deficiency value and normalized mean deficiency value. On row 19, the control corpus of size 475k records the 15.1643 as the least normalized mean value among the various control corpus sizes listed in the above table

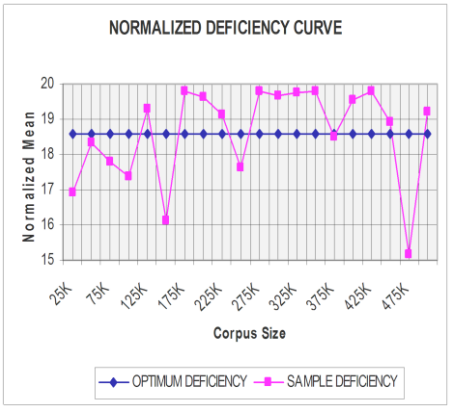

Figure.1. Normalized Deficiency

Normalized Sampling deficiency is shown in the fig 4.4 wherein, the x axis represents the control corpus size whose values range from 25k to 500k and the y axis represents the values of Normalized sampling deficiency. It is observed that most of the values are closer to the optimum value. Here the optimum value is 18.6033 whereas the maximum normalized sampling deficiency value is 19.8003 and the minimum normalized sampling deficiency value is 15.1643.

3. RESULTS AND DISCUSSION

Mean Sample Size Weighted Failure Rate: In this section, we define a new metric called, Mean Sample Size Weighted Failure Rate, as follows: Mean Sample Size Weighted Failure Rate (MSSWFR) is given by

$$\text{MSSWFR} = \frac{\sum_{i=1}^{n}(\mu f_i * S_i)}{\sum_{i=1}^{n} S_i}$$

Where,

S_i = Control corpus size for a given i,

μf_i = Average Failure rate for a given control corpus i.

It may be noted that when $S_i = S$,

MSSWFR = $\sum_{i=1}^{n}(\mu f_n)/n$

The following values are need to be considered from the table 4

Total corpus size of all the sample population = 5250000

Total Sample Mean of all the sample population = 366.3333

Total Corpus Mean value of all the sample population = 101536111

Average Sample Mean of all the sample population = 366.3333 / 20 = 18.3166

$$\text{Average error rate of the corpus population} = \frac{\text{Total corpus mean}}{\text{Total corpus size}} = \frac{101536111}{5250000} = 19.3402$$

Total Corpus Mean

Since the difference between the Corpus Mean and Sample Mean is fairly minimal, we conclude that the error rate is within the acceptable region.

Table.4. **Sampling Mean Error Analysis**

S.No	Control Corpus (A)	Sample Mean (B)	Corpus Mean (C = A * B)
1	25K	14.5556	363889
2	50K	15.6667	783333
3	75K	15.2222	1141667
4	100K	14.8889	1488889
5	125K	16.7778	2097222
6	150K	14.0000	2100000
7	175K	18.3333	3208333
8	200K	19.2222	3844444
9	225K	20.1111	4525000
10	250K	21.5556	5388889
11	275K	18.6667	5133333
12	300K	19.1111	5733333
13	325K	17.7778	5777778
14	350K	18.4444	6455556
15	375K	20.7778	7791667
16	400K	19.4444	7777778
17	425K	18.2222	7744444
18	450K	20.3333	9150000
19	475K	23.2222	11030556
20	500K	20.0000	10000000
Total	**5250K**	**366.3333**	**101536111**

The calculated value of MSSWFR equals 19.3402 which is close to the non-weighted mean failure rate of 18.3166. This clearly indicates that the KanchiCrawler algorithm works efficiently by classifying relevant URLs across the control corpus of varying sizes.

4. CONCLUSION

In this paper, a novel approach is proposed to estimate the error / failure rate of the focused web crawler during the classification stage. It also introduces the new metric Mean Sample Size Weighted Failure Rate (MSSWFR). MSSWFR value is computed and compared with the non-weighted mean failure rate value to estimate the error. As a consequence, we are justified in using Mean and Standard Deviation to develop further analysis and modeling with a reasonable degree of confidence.

REFERENCES

Abiteboul, Serge, Mihai Preda, and Gregory Cobena, Adaptive on-line page importance computation, Proceedings of the 12th international conference on World Wide Web. ACM, 2003.

Avraam, Ioannis, and Ioannis Anagnostopoulos,A comparison over focused web crawling strategies, Informatics (PCI), 2011 15th Panhellenic Conference on. IEEE, 2011.

Baroni, Marco, and Motoko Ueyama, Building general-and special-purpose corpora by web crawling, Proceedings of the 13th NIJL international symposium, language corpora, Their compilation and application, 2006.

Chakrabarti S, van den Berg M, and Dom B, Focused crawling, a new approach to topic-specific Web resource discovery, Computer Networks (Amsterdam, Netherlands), 31, 1999, 1623–1640.

Hong-Wei Hao, Cui-Xia Mu, Xu-Cheng Yin, Shen Li, Zhi-Bin Wang, An Improved Topic Relevance Algorithm for Focused Crawling 978-1-4577-0653- 0/11, 2011

Hozo, Stela, Benjamin Djulbegovic, and Iztok Hozo.,Estimating the mean and variance from the median, range, and the size of a sample, BMC medical research methodology 5(1), 2005,1.

Kilgarriff, Adam, and Gregory Grefenstette, Introduction to the special issue on the web as corpus, Computational linguistics 29(3), 2003, 333-347.

Lilliefors, Hubert W,On the Kolmogorov-Smirnov test for normality with mean and variance unknown, Journal of the American Statistical Association, 62(318), 1967, 399-402.

McKay, Michael D, Richard J Beckman, and William J Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 42(1), 2000, 55-61.

Moore, David S. The basic practice of statistics, New York, WH Freeman, 2, 2007

Safran, Mejdl S, Abdullah Althagafi, and Dunren Che., Improving relevance prediction for focused Web crawlers, Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on. IEEE, 2012.

Soumen Chakrabarti, Mining the Web, from Chapters 2, 3, giga-pedia, 17-77.

Wei-jiang, Li, A New Algorithm of Topical Crawler, Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on, IEEE, 1, 2009.

Zhou L, Bing, A distributed vertical crawler using crawling-period based strategy, Future Computer and Communication (ICFCC), 2010 2nd International Conference on, IEEE, 1, 2010.