ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(35), DOI: 10.17485/ijst/2016/v9i35/101809, September 2016* ISSN (Online) : 0974-5645

An M/M/1 Based Modeling Approach for the Web Crawled Data

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

[1]Research Scholar, Department of CSA, SCSVMV University, Enathur, Tamil Nadu, India; prof.lrajesh@gmail.com
[2]Professor, Department of MCA, St. Joseph's College of Engineering, Chennai, India; drvshanthi@yahoo.co.in
[3]Professor, Department of Computer Science & Engineering, RGMCET, Andhra Pradesh, India; vlnsimhan60@gmail.com

Abstract

Objectives: To develop a suitable model to study the behavior of web crawled dataset and perform simulation on the modeled data for better understanding of the system **Methods/Statistical Analysis:** M/M/1 model is a variation of Single Birth Single Death (SBSD) model which is applied to study the behavior of web crawled dataset for the Classification Problem. KanchiCrawler, a stylized focused web crawler is implemented to collect the data for this application. The size of the corpora (Population) is 500k. Control corpus (sample) can be drawn from the corpora based on enforcing certain pre-determined conditions. Findings: A 20-state model starting with an initial test corpus of 25k and then by gradually increasing with an increment of 25k up to 500k is developed. This is achieved through the computation of Forward State Transition Probability and Reverse State Transition Probability for the respective states. This model provides fairly good results by testing the algorithmic efficiency of a KanchiCrawler and to model the web crawled dataset for the classification problem. **Applications:** M/M/1 models are tractable and often used to model various operations of nature. In most situations where large numbers are involved, M/M/1 model are statistically stable and reflective of reality.

Keywords: Dataset Modeling, KanchiCrawler, M/M/1 Model, State Transition Probability

1. Introduction

Web Crawler are programs used to download documents from the internet. It starts downloading the documents with the initial set of URLs called seed URLs. Web Crawler stops crawling either downloading the entire URL or it reaches the threshold limit. The collection of URLs which is downloaded by the Web Crawler forms the data set. The main advantage of using topical crawler[1] over generalized crawler is that the former concerns only about the particular topic of search. Various analyses is performed on the Web Crawler Data set. KanchiCrawler, a stylized focused web crawler, is used to gather the data for this application. In order to study and understand the behaviour of any dataset, they have to be modeled and simulated. A model is an approximation of the system and modeling is generally developed through the profiling of multiple parameters. A model allows a researcher to understand, characterize the system and experiment with the impact of *what-if* conditions. Models can be micro model or macro model or the combination of thereof. Micro models model a small system or event while macro models model a large system or set of events. Micro models therefore can be combined in order to generate macro models. Micro-model is also known as fine-grained model. Models can also incorporate feedbacks, stability analysis, performance factors and their measurements and consider many other such issues. A model therefore is an abstraction of the system, but not the system itself. Many models are developed after observing the behaviour of an actual system. However, in some cases one can only observe the black box behaviour of the system and in such a case modeling helps to understand the white box nature of the system[2]. This paper presents a novel approach to model the web crawled data using Single Birth Single Death Model. KanchiCrawler, a stylized focused web crawler is implemented to collect the data for this application. Single Birth Single Death (SBSD) modeling

* *Author for correspondence*

is implemented for both successful as well as the Failure case. Various transition probabilities of Forward state and Reverse state were also computed. The rest of the paper is organized as follows: section 2 of the paper provides a brief overview of modeling techniques relevant to this paper, while section 3 details the application of Single Birth Single Death Model for an web crawled dataset. Section 4 offers the results and related discussion, while the conclusion summarizes the paper and offers pointers for further work in this arena.

2. Modeling Techniques – A Review

There are many tools and techniques available to perform a modeling and some of them include but not limited to the following types.

2.1 Petri net Based Modeling

Petri net Based Modeling was first introduced by Carl Adam Petri in1962. Petri net is a diagrammatic tool to model concurrency and synchronization in distributed systems. They have been used to model and analyze several types of processes including protocols, manufacturing systems, and business processes[3]. This modeling technique is somewhat similar to State Transition Diagrams. It is used as a visual communication aid to model the system behaviour. Petri nets[4-6] are a well-founded process modeling techniques that have formal semantics. Petri net is based on strong mathematical foundation wherein events/action and the state of the system are modeled as executable/firable directed acyclic graph. The firing represents an occurrence of the event or an action taken. The theory of Petri net[7] is quite extensive and has been used to model computing systems, physical systems, social systems and even the human systems. This modeling technique consists of three types of components: *places* (circles), *transitions* (rectangles) and *arcs* (arrows). Places represent possible states of the system; Transitions are events or actions which cause the change of state and every arc simply connects a place with a transition or a transition with a place. In modeling, the firing of a transition simulates the occurrence of that event[8]. An event can take place only if all of the conditions for its execution have been met; that is, the transition can be fired only if it is enabled.

2.2 Stochastic Modeling[9]

Stochastic Modeling concerns the modeling of action and states of a system using probabilistic/stochastic distributions. Stochastic model permits both discrete and continuous variations in the modern parameters to be considered and the nature of the system analyzed accordingly.

2.3 Queuing Model

Queuing model concerns the applications of the queuing theory for modeling the systems[10]. Queuing theory has been well-developed with the use of a number of continuous and discrete time domain distributions, which facilitates many variations of the models to be evaluated[11].

2.4 Finite element Models and Piecewise Models[12]

Finite element models and Piecewise models finite element model[13] (mostly used in structural analysis concerns the modeling of a micro system under the particular input output environment variables, while piecewise models deal with modeling only the part of the system under certain conditions; such models can be both linear and non-linear. Many other modeling techniques exists which include dynamic modeling (model is modified dynamically), real time models (modeling parameters vary over time and hence the systems behavior overall themselves) and enumerative models (typically used in sociology, psychology and anthropology)

3. Proposed Approach

3.1 Single Birth and Single Death Model

SBSD is queuing theory based modeling approach. SBSD Model relates to a system with a population x, wherein it is mandated that only a single birth or a single death occurs at any given time. Depending on the rate of birth (traditionally called λ) and rate of death traditionally called μ) the population of the system can grow (and explode) or be stable or deplete to the point of no return. However in general, the population in the system exhibits oscillatory behavior of expansion and contraction. While this model is useful to study the behavior of small human subsystems (sociologically cohesive units in a village or a town), it is not reflective of the nature. In the latter case, Multiple Birth and Multiple Death (MBMD) models can

be used. Inherently these models assume the distributions such as the Normal distribution or Poisson distribution. In general, one can assume General Birth and General Death (GBGD), but the solution then becomes intractable. In this paper, only SBSD model is assumed, since it gives a simple and tractable way to model the operations of sampling efficiency and sampling deficiency. Using this model, we have obtained fairly good results. M/M/1 model is a variation of SBSD model, wherein the birth and death rates are assumed to follow normal distribution. Further the population is assumed to be unity (i.e., predefined number or an entity) to start with. M/M/1 models are tractable and often used to model various operations of nature. In most situations where large numbers are involved, M/M/1 model are statistically stable and reflective of reality. Therefore in this paper, we shall adapt M/M/1 model to model sampling efficiency and sampling deficiency. We have assumed that a SBSD model for the operation of our system is shown in Figure 1, wherein the various Forward State Transition Probabilities (FSTPs) and Reverse State Transition Probabilities (RSTPs) can be calculated. The individual values for the FSTP and RSTP probabilities are obtained from Table 1. Figure 1 and 2 provide the state diagram for the M/M/1 model for our

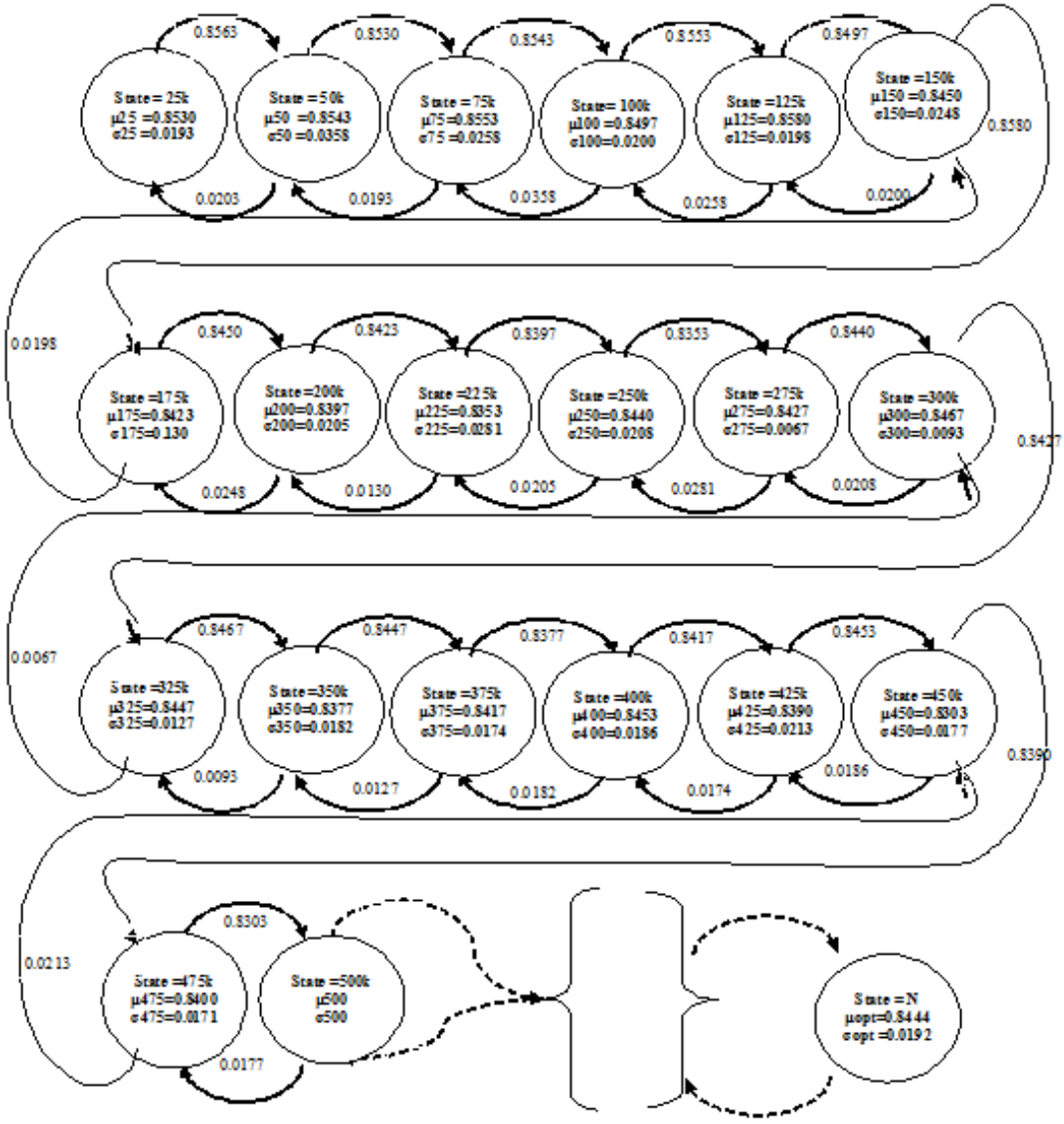

Figure 1. SBSD Model for success.

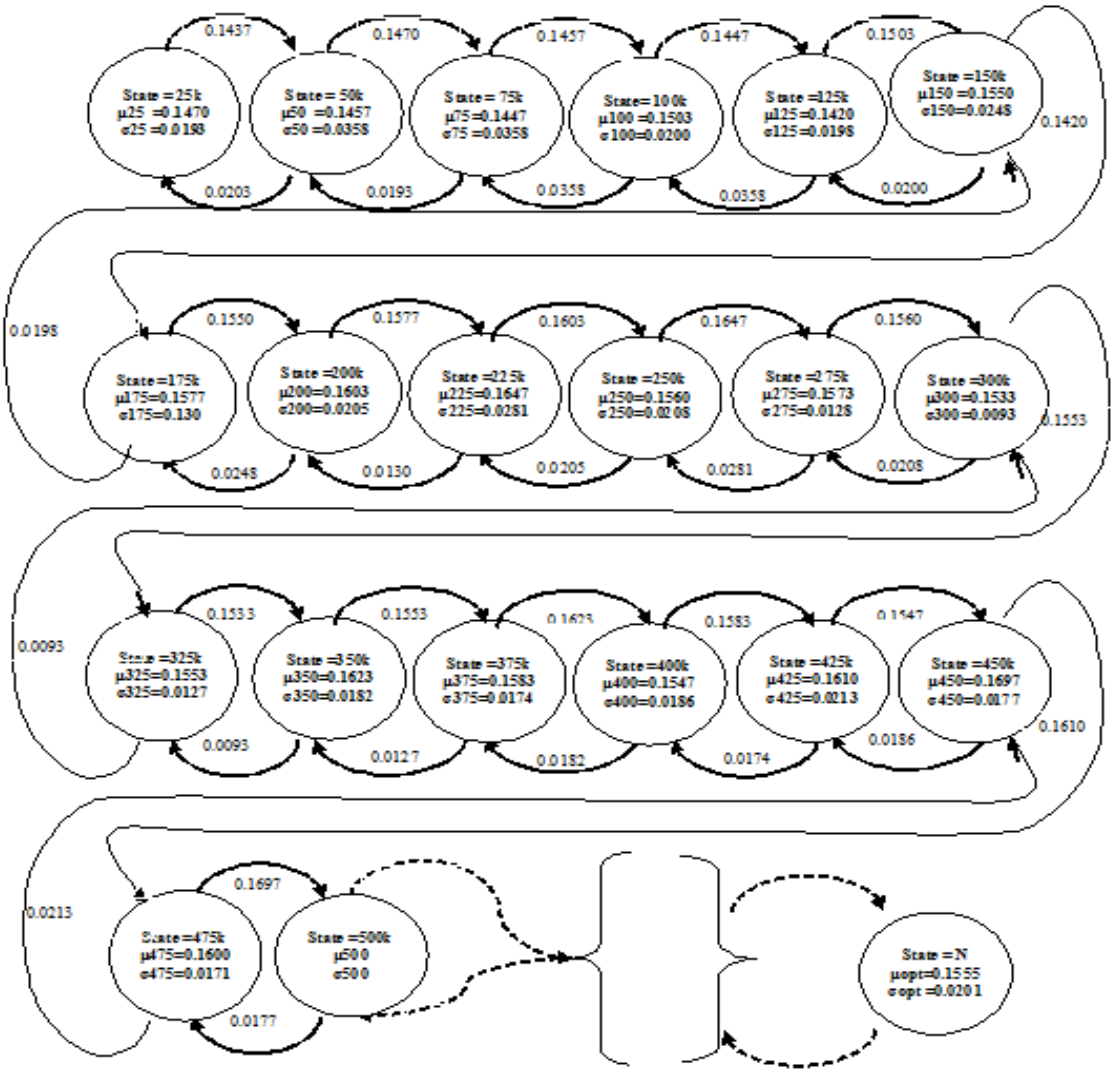

Figure 2. SBSD Model for Failure.

system, wherein the steady state values of the parameters of the model can be calculated as the state size increases. The computed steady state values for success are given:

N= (Practical Infinity)

$Mean_{optimum}$ = 0.8444

$SD_{optimum}$ =0.0192

Note: Forward and Reverse State Transition Probabilities for success

Note: Forward and Reverse State Transition Probabilities for Failure

The steady state distribution can then be modeled as follows:

- With the available information, a simple model of type SBSD[1] for both the cases of success and failure has been developed.
- The analysis is computed with an initial test corpus of 25k and then by gradually increasing with an increment of 25k.
- A 20-state model - starting from 25k to 500k – is developed thru' the computation of FSTPs and RSTPs for the respective states.
- The FSTP for a given state S is computed by using the following formula:

$P_{FSTP} \mid P_{FSTP-1}$ = Sampling Efficiency at (s)

Where, Sampling Efficiency is the total number of identified instrumented URLs to the total number of instrumented URLs. The instrumented URLs are the

known set of URLs which are added/instrumented to the control corpus in order to test the efficiency of the crawler algorithm. Control corpus (sample) can be drawn from the corpora (population) based on enforcing certain pre-determined conditions. We compute the Reverse State Transition Probability (RSTP) for a given state S by using the formula:

$\mathbf{P_{RSTP} \mid P_{RSTP+1} = Sampling\ Efficiency\ at\ (s)}$

The computed steady state values for failure are given below:

N= (Practical Infinity)

$Mean_{optimum} = 0.1555$

$SD_{optimum} = 0.0201$

Table 1. Forward State Transition Probability for success

Sl. No.	State	Mean Sample Efficiency	SD Sampling Efficiency
1	P50/25	0.8530	0.0193
2	P75/50	0.8543	0.0358
3	P100/75	0.8553	0.0258
4	P125/100	0.8497	0.0200
5	P150/125	0.8580	0.0198
6	P175/150	0.8450	0.0248
7	**P200/175**	**0.8423**	**0.0130**
8	P225/200	0.8397	0.0205
9	P250/225	0.8353	0.0281
10	P275/250	0.8440	0.0208
11	P300/275	0.8427	0.0067
12	P325/300	0.8467	0.0093
13	P350/325	0.8447	0.0127
14	P375/350	0.8377	0.0182
15	P400/P375	0.8417	0.0174
16	P425/400	0.8453	0.0186
17	P450/425	0.8390	0.0213
18	P475/450	0.8303	0.0177
19	P500/475	0.8400	0.0171

4. Results and Discussions

Let us consider the Table 1 which details forward state transition probability for success. The table lists the values of state position, Mean Sample Efficiency and Standard Deviation Sample Efficiency. In Row 5, for the state 150/125, the sampling efficiency has the maximum value of 0.8580. The various state values of sampling efficiency are comparatively closer to the optimum value which is computed as 0.8444. Let us consider the Table 2 which details Reverse state transition probability for success. The Table lists the values of state position, Mean Sample Efficiency and Standard Deviation Sample Efficiency. In Row 8, P300/325, the value 0.0067, is the least Standard Deviation Sampling efficiency value. The optimum value for the Standard Deviation Sampling Efficiency is computed as 0.0193.Let us consider the Table 3 which details Forward state transition probability for failures. The table lists the values of state position, Mean Sample Deficiency and Standard Deviation Sample Deficiency. In Row 5, for the state 150/125, the sampling deficiency has the minimum value of 0.1420. The various state values of sampling deficiency are comparatively closer to the optimum value which is computed as 0.1555. Let us consider the Table 4 which details Reverse state transition probability for failures. The table lists the values of state position, Mean Sample Deficiency and Standard Deviation Sample Deficiency.

Table 2. Reverse State Transition Probability for success

Sl. No.	State	Mean Sample Efficiency	SD Sampling Efficiency
1	P475/500	0.8303	0.0177
2	P450/475	0.8390	0.0213
3	P425/450	0.8453	0.0186
4	P400/425	0.8417	0.0174
5	P375/400	0.8377	0.0182
6	P350/375	0.8447	0.0127
7	P325/350	0.8467	0.0093
8	P300/325	0.8427	0.0067
9	P275/300	0.8440	0.0208
10	P250/275	0.8353	0.0281
11	P225/250	0.8397	0.0205
12	P200/225	0.8423	0.0130
13	P175/200	0.8450	0.0248
14	P150/175	0.8580	0.0198
15	P125/150	0.8497	0.0200
16	P100/125	0.8553	0.0258
17	P75/100	0.8543	0.0358
18	P50/75	0.8530	0.0193
19	P25/50	0.8563	0.0203

Table 3. Forward State Transition Probability for failure

Sl. No.	State	Mean Sample Deficiency	SD Sampling Deficiency
1	P50/25	0.1470	0.0193
2	P75/50	0.1457	0.0358
3	P100/75	0.1447	0.0358
4	P125/100	0.1503	0.0200
5	P150/125	0.1420	0.0198
6	P175/150	0.1550	0.0248
7	P200/175	0.1577	0.0130
8	P225/200	0.1603	0.0205
9	P250/225	0.1647	0.0281
10	P275/250	0.1560	0.0208
11	P300/275	0.1573	0.0128
12	P325/300	0.1533	0.0093
13	P350/325	0.1553	0.0127
14	P375/350	0.1623	0.0182
15	P400/P375	0.1583	0.0174
16	P425/400	0.1547	0.0186
17	P450/425	0.1610	0.0213
18	P475/450	0.1697	0.0177
19	P500/475	0.1600	0.0171

Table 4. Reverse State Transition Probability for failure

Sl. No.	State	Mean Sample Deficiency	SD Sampling Deficiency
1	P475/500	0.1697	0.0177
2	P450/475	0.1610	0.0213
3	P425/450	0.1547	0.0186
4	P400/425	0.1583	0.0174
5	P375/400	0.1623	0.0182
6	P350/375	0.1553	0.0127
7	P325/350	0.1533	0.0093
8	P300/325	0.1573	0.0128
9	P275/300	0.1560	0.0208
10	P250/275	0.1647	0.0281
11	P225/250	0.1603	0.0205
12	P200/225	0.1577	0.0130
13	P175/200	0.1550	0.0248
14	P150/175	0.1420	0.0198
15	P125/150	0.1503	0.0200
16	P100/125	0.1447	0.0358
17	P75/100	0.1457	0.0358
18	P50/75	0.1470	0.0193
19	P25/50	0.1437	0.0203

5. Conclusions

In this paper, a novel approach is proposed to build M/M/1 based model for the web crawler dataset. In this paper, only SBSD model is assumed, since it gives a simple and tractable way to model the operations of sampling efficiency and sampling deficiency. A 20-state model - starting from 25k to 500k – is developed thru' the computation of FSTPs and RSTPs for the respective states. Using this model, we have obtained fairly good results and hence we left the other advanced modeling as a pointer for future research.

6. References

1. Amandeep Verma, Amandeep Kaur Gahier. Topic Modeling of E-News in Punjabi. Indian Journal of Science and Technology. 2015 Oct; 8(27):1-10.
2. Peyman Salah, Seyed Siavash Karimi Madahi, Hassan Feshki Farahani, Ali Asghar Ghadimi. A New Method to Calculate Residential Consumer's Consumption Using Computer Modeling. Indian Journal of Science and Technology. 2012 May; 5(5):1-5.
3. Van Der Aalst WM. Process-oriented architectures for electronic commerce and inter organizational workflow. Information systems. 1999 Dec; 24(9):639-71.
4. Peterson JL. Petri net theory and the modeling of systems. Prentice Hall PTR Upper Saddle River, NJ, USA. 1981.
5. Shamim Yousefi, Samad Najjar Ghabel, Leyli Mohammad Khanli. Modeling Causal Consistency in a Distributed Shared Memory using Hierarchical Colored Petri Net. Indian Journal of Science and Technology. 2015 Dec; 8(33):1-7.
6. Bahman AR, Alialhosseini E. Modeling of Component Diagrams Using Petri Nets. Indian Journal of Science and Technology. 2010 Dec; 3(12):1-11.
7. Reisig W. Petri Nets: An Introduction, volume 4 of Monographs in Theoretical Computer Science. Springer. 1985 May.
8. Johnsonbaugh R, Murata T. Petri nets and marked graphs-mathematical models of concurrent computation. The American Mathematical Monthly. 1982 Oct; 89(8):552-66.
9. Sawyer H, Kauffman MJ. Stopover ecology of a migratory ungulate. Journal of Animal Ecology. 2011 Sep; 80(5):1078-87.
10. Shankar, Natarajan. Symbolic Analysis of Transition Systems? ASM '00 Proceedings of the International Workshop on Abstract State Machines, Theory and Applications. 2000, p.287-302.
11. Larson RC. A hypercube queuing model for facility location and redistricting in urban emergency services. Computers and Operations Research. 1974 Mar; 1(1):67-95.
12. Leenaerts D, Van Bokhoven WM. Piecewise linear modeling and analysis. Kluwer Academic Publishers Norwell, MA, USA. 1998.
13. Ali Kiani, Mohsen Izadinia. An Overview on Effects of Geometric Parameters in Column Connection Behavior via Finite Element Method. Indian Journal of Science and Technology. 2015 Oct; 8(28):1-7.