Computer Networks 55 (2011) 1880–1893

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Optimal threshold control by the robots of web search engines with obsolescence of documents

Konstantin Avrachenkov [a,*], Alexander Dudin [b], Valentina Klimenok [b], Philippe Nain [a], Olga Semenova [b]

[a] INRIA Sophia Antipolis, 2004 Route des Lucioles, B.P. 93, Sophia Antipolis, 06902, France
[b] Belarusian State University, 4 Independence Ave., Minsk-30, Belarus

ARTICLE INFO

Article history:
Received 7 October 2009
Received in revised form 7 December 2010
Accepted 25 January 2011
Available online 11 February 2011

Keywords:
Search engine
Crawler
Queueing system

ABSTRACT

A typical web search engine consists of three principal parts: crawling engine, indexing engine, and searching engine. The present work aims to optimize the performance of the crawling engine. The crawling engine finds new web pages and updates web pages existing in the database of the web search engine. The crawling engine has several robots collecting information from the Internet. We first calculate various performance measures of the system (e.g., probability of arbitrary page loss due to the buffer overflow, probability of starvation of the system, the average time waiting in the buffer). Intuitively, we would like to avoid system starvation and at the same time to minimize the information loss. We formulate the problem as a multi-criteria optimization problem and attributing a weight to each criterion. We solve it in the class of threshold policies. We consider a very general web page arrival process modeled by Batch Marked Markov Arrival Process and a very general service time modeled by Phase-type distribution. The model has been applied to the performance evaluation and optimization of the crawler designed by INRIA Maestro team in the framework of the RIAM INRIA-Canon research project.

1. Introduction

The problem of control by the robots (crawlers) that traverse the web and bring web pages to the indexing engine that updates the data base of a web search engine is formulated and analyzed in [13]. This problem is formulated in [13] as the controlled queueing system. The system has a single server with the exponential service time distribution, finite buffer of capacity $K-1$, $K \geqslant 2$. There are N available robots and each of these robots, when activated, brings pages to the server in a Poisson stream at fixed rate. These N stationary Poisson processes are mutually independent and independent of service times.

The number of active robots may be modified at any arrival or departure event. When an arrival occurs, one or several active robots can be de-activated at once. When a departure occurs the controller may decide either to activate one or several available non-active robots. Of course, the controller can choose to do nothing (i.e. the number of active robots remains the same).

In [13], the problem of finding a policy that minimizes a weighted sum of the loss rate and starvation probability (probability of the empty system) is considered. It is solved by means of the tools of the Markov Decision Processes theory.

The following possible generalizations of the model, which are certainly worthwhile analyzing, are mentioned in [13]:

- More general input processes, e.g., an *MMPP* (*Markov Modulated Poisson Process*) should be considered so as

[*] Corresponding author.
E-mail addresses: K.Avrachenkov@sophia.inria.fr (K. Avrachenkov), dudin@bsu.by (A. Dudin), klimenok@bsu.by (V. Klimenok), nain@sophia.inria.fr (P. Nain), olgasmnv@gmail.com (O. Semenova).

1389-1286/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2011.01.013

to reflect more accurately "traveling times" of robots in the network;

- Because of the obsolescence of stored documents issue, the waiting time should be bounded, even if the buffer size is effectively infinite;
- Other cost functions could be investigated, for instance, cost functions including response times.

In this paper, we made all the mentioned and some further generalizations.

We assume that, under the fixed number of currently active robots, the arrival process is of the *BMAP* type. The *BMAP* is a more general process comparing to the *MMPP* and allows delivering a batch of web pages to be indexed while the *MMPP* assumes that the pages are delivered one-by-one. It is very typical for a computer system to operate in batch mode. Also BMAP allows delivering pages at the phase change time moments.

We assume that the service time distribution is of the *PH* (Phase) type which is much more general comparing to the exponential distribution assumed in [13]. The class of phase type distributions is dense in the field of all positive-valued distributions and practically we can deal with any real distribution [2].

Since web pages can become obsolete, we bound the waiting time stochastically. Waiting time of each web page in a buffer is restricted by a random variable having *PH* distribution identical and mutually independent for all web pages. The phase type distribution has been used to model obsolescence times for instance in [15].

We suppose that the cost function can have a more general form than in [13] and include the cost of the obsolescence and the response time.

In the next section we formulate the model and the optimization problem. Section 3 contains the steady-state analysis of the multi-dimensional Markov chain which defines dynamics of the system under the fixed values of the parameters defining the control strategy. In Section 4 the main performance measures of the system are computed. In Section 5 the conditional sojourn time distributions are calculated. In Section 6 a particular case of ordinary arrivals is discussed. In Section 7 the theoretical results are illustrated by numerical examples. In particular, the mathematical model is applied to the performance evaluation and optimization of the web crawler designed by INRIA Maestro team in the framework of the RIAM INRIA-Canon research project. Section 8 concludes the paper.

2. Mathematical Model

We consider a single server system with the finite buffer of capacity $K-1$, $K \geqslant 2$. So, the total number of web pages which can stay in the system is restricted by the number K. Web pages are served by a server in the order of their arrivals.

Service times of web pages are independent identically distributed random variables having *PH* distribution with irreducible representation $(\boldsymbol{\beta}, S)$. It means the following. Service of a web page is defined as a time until the continuous-time Markov chain m_t, $t \geqslant 0$, having the states $(1,\ldots,M)$ as the transient and state 0 as absorbing one reaches the absorbing state. An initial state of the chain is selected in a random way, according to the probability distribution defined by the row-vector $(\boldsymbol{\beta},0)$, where $\boldsymbol{\beta}$ is the stochastic row vector of dimension M. Transitions of the Markov chain m_t, $t \geqslant 0$, are described by the generator $\begin{pmatrix} S & S_0 \\ 0 & 0 \end{pmatrix}$ where the matrix S is a sub-generator and the column vector S_0 is defined by $S_0 = -S\mathbf{e}_M$ and has all non-negative and at least one positive components, $\mathbf{e}_M$ is the column vector of dimension M consisting of all 1's. The average service time b_1 is given by $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}_M$. For more details about the *PH* type distribution, its properties, special cases and applications see [11,12].

Web pages can be delivered into the system by N available robots. The number of active robots varies in the set $\{1,\ldots,N\}$. We assume that the process of web pages delivering under l, $l=\overline{1,N}$, active robots is described as follows. Let ν_t, $t \geqslant 0$, be an irreducible continuous time Markov chain having finite state space $\{0,1,\ldots,W\}$. Sojourn time of the chain ν_t, $t \geqslant 0$, in the state ν has exponential distribution with a parameter $\lambda_\nu^{(l)}$. After this time expires, with probability $p_0^{(l)}(\nu,\nu')$ the chain jumps into the state ν' without generation of web pages and with probability $p_k^{(l)}(\nu,\nu')$ the chain jumps into the state ν' and a batch consisting of k web pages is generated, $k \geqslant 1$. The introduced probabilities satisfy conditions:

$$p_0^{(l)}(\nu,\nu)=0, \quad \sum_{k=1}^{\infty}\sum_{\nu'=0}^{W} p_k^{(l)}(\nu,\nu') + \sum_{\nu'=0}^{W} p_0^{(l)}(\nu,\nu') = 1,$$
$$\nu=\overline{0,W}, \quad l=\overline{1,N}.$$

The parameters defining this flow are kept in the square matrices $\mathcal{D}_k^{(l)}$, $k \geqslant 0$, $l=\overline{1,N}$, of size $\overline{W}=W+1$ defined by their entries:

$$\left(\mathcal{D}_0^{(l)}\right)_{\nu,\nu} = -\lambda_\nu^{(l)}, \quad \left(\mathcal{D}_0^{(l)}\right)_{\nu,\nu'} = \lambda_\nu^{(l)} p_0(\nu,\nu'),$$
$$\left(\mathcal{D}_k^{(l)}\right)_{\nu,\nu'} = \lambda_\nu^{(l)} p_k^{(l)}(\nu,\nu'), \nu,\nu'=\overline{0,W}, \quad k \geqslant 1, \quad l=\overline{1,N}. \tag{1}$$

Denote the generating function

$$\mathcal{D}^{(l)}(z) = \sum_{k=0}^{\infty} \mathcal{D}_k^{(l)} z^k, \quad |z| \leqslant 1.$$

The matrix $\mathcal{D}^{(l)}(1)$ is the infinitesimal generator of the process $\nu_t, t \geqslant 0$, under the fixed number l of active robots. The stationary distribution vector $\boldsymbol{\theta}^{(l)}$ of this process satisfies the equations $\boldsymbol{\theta}^{(l)}\mathcal{D}^{(l)}(1) = \mathbf{0}$, $\boldsymbol{\theta}^{(l)}\mathbf{e} = 1$. Here and in the sequel, $\mathbf{0}$ is the zero row vector. The average intensity $\lambda^{(l)}$ (fundamental rate) of the *BMAP* under the fixed number l of active robots is defined by

$$\lambda^{(l)} = \boldsymbol{\theta}^{(l)} \frac{d\mathcal{D}^{(l)}(z)}{dz}|_{z=1}\mathbf{e},$$

and the intensity $\lambda_g^{(l)}$ of group arrivals is defined by

$$\lambda_g^{(l)} = \boldsymbol{\theta}^{(l)}\left(-\mathcal{D}_0^{(l)}\right)\mathbf{e}.$$

The variance $\mathcal{V}^{(l)}$ of intervals between group arrivals is calculated as follows:

$$\mathcal{V}^{(l)} = 2\left(\lambda_g^{(l)}\right)^{-1}\boldsymbol{\theta}^{(l)}\left(-\mathcal{D}_0^{(l)}\right)^{-1}\mathbf{e} - \left(\lambda_g^{(l)}\right)^{-2},$$