Knowledge-Based Systems 149 (2018) 99–109

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Fluid models for call centers with delay announcement and retrials

Miao Yu[a,*], Jiafu Tang[b], Fanwen Kong[a], Chunguang Chang[a]

[a] School of Management, Shenyang Jianzhu University, China
[b] School of Management Science and Engineering, Dongbei University of Finance and Economics, China

ARTICLE INFO

Article history:
Received 7 October 2017
Revised 24 February 2018
Accepted 27 February 2018
Available online 6 March 2018

Keywords:
Predicting and announcing delays
Fluid model
Call center
Abandoned calls
Retrials
Overloaded queues

ABSTRACT

This paper models a call center as a multi-server queue where anticipated delays are announced to customers upon arrival, and customer balking, reneging and retrials are modeled explicitly. The resulting queue with delay announcement is modeled in a stationary setting. We propose a fluid approximation to estimate the possibility of announcing the mean delay distribution and the retrials in the system. This approximation method can overcome some of the computational issues involved with a continuous time Markov chain analysis. The fluid approximation is also validated that it can work well in overloaded systems by performing a comparison between the fluid model and the stochastic model with delay information, and by performing the other comparison between the fluid model and the simulation model. Through a numerical study, this paper demonstrates the significance of delay information in a call center with retrials. In particular, delay announcement is more important in the system with a high retrial probability. Thus, we show how a delay announcement greatly reduces customer reneging and thereby improves customer satisfaction. It is shown that disregarding retrials in call centers with delay information may result in large distortions in the management of call centers.

1. Introduction

Motivated by applications to call centers, there has been great interest in multiserver queues with delay information and customer reactions [3,12]. Announcing delay information to customers can modulate demand by announcing times of high congestion and enhancing satisfaction with inevitable waiting. Some customers are patient enough to wait for the service upon receiving delay information, while others may hang up or abandon the call after waiting for some time. Furthermore, a portion of customers who abandoned the call will redial and try to access the call center. Customers who do not like to wait will abandon or attempt a call several times. The purpose of this paper is to model these features of the queuing system, with customer reactions to delay information and retrial phenomena. The model is a deterministic fluid model, and the system with heavy loads is worth considering.

A queuing system with delay information captures the customer psychology associated with uncertain waiting [21]. In practice, delay announcements can play a distinct role in increasing customer patience by reducing the uncertainty, and affect customers behavior in terms of abandonment (balking and reneging). In the last decade, whether managers or researchers all emphasized the importance of providing delay information, thus a great deal of research achievements emerged. In broad terms, there are three mains areas of research. The first area studies the effect of delay announcements on system dynamics. Specifically, a large number of studies on the impact of delay information have focused on the system performance in the invisible queue [29]. Guo and Zipkin [13] study a model in which customers are provided with delay information and make decisions based on their expected wait times. Aksin et al. [4] combine modeling and empirical analysis in an analysis of delay announcements in a call center. Jouini et al. [19] propose a new framework that uses a news-vendor-like performance criterion to pick the value to announce from the estimated delay distribution. Research has found that informing customers of delays is beneficial regardless of the model used, but the optimal amount of precision in the announcements varies from model to model. Furthermore, the importance of modeling customer responses in the following literature is emphasized [20]. Armony et al. [7] study customer responses to delay information by requiring an equilibrium analysis. Armony and Maglaras [5,6] consider a system where the service manager provides the customers with an estimate of the delay, and a customer may balk or wait based on this information. In the recent work, Yu et al. [31] posit that delay announcements impact customer behavior in a complex way. The second area studies alternative ways of delay estimation in service systems: e.g. Ibrahim and Whitt [16–18]. The third area is mainly focused on customer psychology in waiting situations:

* Corresponding author.
E-mail addresses: yumiao1213@126.com (M. Yu), jftang@mail.neu.edu.cn (J. Tang), kfanwen@163.com (F. Kong), 1523537959@qq.com (C. Chang).

https://doi.org/10.1016/j.knosys.2018.02.040
0950-7051/

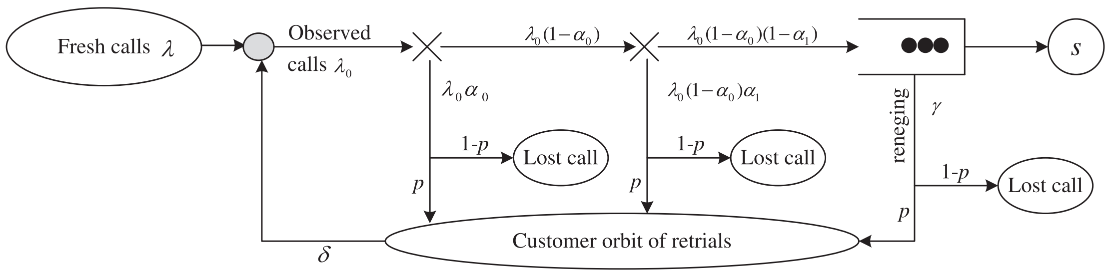

Fig. 1. A call center without delay information, balking, reneging, and retrial.

e.g., see Hui and Tse [15], Munichor and Rafaeli [24], Taylor [27]. This paper research on delay announcements falls into the first main area of research.

In practice, blocked or abandoned calls may be attempted later, that is, customers were unable to enter the system on the first attempt. This type of retrial behavior and its influence is modeled by Haddad et al. [14] and Shin and Moon [26]. Mandelbaum et al. [22] consider multiserver systems with abandonments and retrials and propose a fluid approximation for their analysis. Aguir et al. [1] focus on a steady state Markov chain analysis to study the retrials phenomena The use of queue models with retrial analysis is common in a call center, and estimating redial calls is necessary to optimize the performance of the system. Even though delay information exists in the system, retrials phenomenon should not either be ignored. However, in the context of prediction of delay and estimating retrial rate, the analysis becomes more difficult because we must take into account the description of the system in addition to delay information given to each waiting customer and each retrying customer after abandoning the queue.

Approximations, in particular a fluid approximation, perform very well for these complex systems. The fluid approximation utilizes the key characteristics of the arrival process and the service and patience distributions [8,23,28]. Hence, we develop a deterministic fluid approximation for the M/M/S+M model with a retrial queue and endogenized customer reactions to delay announcements. In the previous studies, this similar model was formulated in Aguir et al. [2]. They propose a fluid model to approximate the queue length process, which tends to be accurate for large overloaded systems. Another similar method was proposed by Ding et al. [9], they make use of fluid approximate to model the customer redial and reconnect behaviors in call centers. However, delay estimates to customers and subsequent reactions of customers are not considered in both models. In our paper, we pay great attention to revealing the impact of delay information on customer reactions and call center performance based on Aguir's researches [1,2].

To the best of our knowledge, this is the first generalization of the fluid model for predicting delays with customer reactions to incorporate retrials. In particular, previous studies usually ignored the role of customer reactions. Therefore, in predicting delays for arriving customers, this paper builds on previous analytical studies from the work of Jouini et al. [20]. They analyzed a call center with impatient customers, in particular, derived the balking and reneging behavior's influence on performances upon hearing the delay announcement. Based on their research achievements, we further explore the case when balking or reneging customers may call back. The model of single-class call centers will not be directly applicable to many current call centers; nevertheless, we think these results and analyses could provide useful and new insight.

The rest of this paper is organized as follows. In the next section, we specify the queuing model in more detail with delay information and retrials. In Section 3, we show how the fluid approximation can be exploited and present the Monte Carlo algorithm to solve the relevant equations. In Section 4, we make numerical comparisons between the fluid approximation and stochastic model in Jouini et al. [20] with only delay information, then make a comparison between the fluid approximation and simulation with delay information and retrial. In Section 5, we show the significance of the retrials and delay information, mainly illustrate the impact of delay information on the system with retrials. In addition, it is illustrated how managers can estimate the first attempt calls from the total observed call arrivals using the model with delay information and the impact of retrials on the management of call center. In Section 6, we provide some concluding remarks and directions for future research.

2. The model

In this section, two cases are considered in a stationary setting: one case where no delay announcement is given to customers, and the other case is based on the process with delay announcement, where customers are informed about their anticipated delay at the time of arrival and customer patience as a function of delay announcement is formulated.

Consider a call center without delay announcement, where the policy for the queue is FCFS (First-Come-First-Served) and abandonment is not allowed once a customer starts service. First attempt customers arrive according to a Poisson process with rate λ. There are s homogeneous servers. The service times are independent and exponentially distributed with mean $1/\mu$. For these customers, assume that the initial random patience time of customers T is random, independent, identically distributed and under a given continuous distribution, given that these patience times are exponentially distributed with parameter γ. The detailed process of the case without delay information is shown in Fig. 1. Customers who cannot access the service at the time of arrival may balk immediately with probability α_0, which of this probability is attributed to ambiguity averse customers. Then there is a probability α_1 who chooses to balk too, and this probability is attributed to uncertainty for delay information. We assume there is no time difference between the two balking decisions, and indirect evidence for this assumption can be found in Pazgal and Radas [25]. These balking customers will both call back with probability p after an exponential distribution of time with the rate δ. This is not necessarily in practice and p may not be a constant probability for subsequent reattempts, but this will be assumed herein for tractability purposes. Customers who join the queue (with probability $(1-\alpha_0)(1-\alpha_1)$) renege if they cannot access the servers within a delay that is exponentially distributed with rate γ. Reneging customers are assumed to have the same retrial behavior as balking customers, and they also call back with probability p and after an exponential distribution of time with rate δ. Aguir et al. [2] supported the assumption that the retrial probability for both balking and reneging customers is approximately equal. This paper refers to the pooling of customers that are waiting to repeat their call as the orbit. Under the above assumptions, the time that a cus-