Автоподбор soft timeout

qkrorlqr@

8 июня 2017 г.

1 Модель

1.1 Цель

Хотим, чтобы ответ источника как можно чаще укладывался в определенное время при заданном ограничении на дополнительную нагрузку. Можем задать один параллельный запрос через некоторое время после исходного. Это некоторое время назовем soft timeout и будем подбирать.

1.2 Предположения

Считаем, что сервис не деградирует при увеличении нагрузки в пределах этого ограничения. Считаем, что времена ответа сервиса на два отдельных запроса независимы (даже если тело запроса одно и то же).

1.3 Обозначения

α - максимальная дополнительная нагрузка (доля от основной нагрузки)
T - время ответа сервиса на отдельно взятый запрос (случайная величина)
t_0 - оптимизируемый квантиль времени ответа (хотим, чтобы источник как можно чаще укладывался в t_0)
s - soft timeout (время с момента отправки исходного запроса, после которого отправляем параллельный запрос)

1.4 Некоторые выкладки

Итого, хотим найти $s^* = argmax_s P(T < t_0 \mid s)$. Путем нехитрых выкладок получаем:

$$P(T < t_0 \mid s) = P(T < t_0) + P(T \geq t_0) \cdot P(T < t_0 - s) \cdot min(\frac{\alpha}{P(T \geq s)}, 1) \tag{1}$$

Выкидываем все, что не зависит от s:

$$s^* = argmax_s \left\{ P(T < t_0 - s) \cdot min(\frac{\alpha}{(1 - P(T < s))}, 1) \right\} \tag{2}$$

В терминах функции распределения T получаем:

$$s^* = argmax_s \left\{ F_T(t_0 - s) \cdot min(\frac{\alpha}{(1 - F_T(s))}, 1) \right\} \tag{3}$$

Рассмотрим 2 региона:
$s > quant_{F_T}(1 - \alpha)$ - здесь имеем $s_r^* = argmax_s \{F_T(t_0 - s)\}$
$s \leq quant_{F_T}(1 - \alpha)$ - здесь имеем $s_l^* = argmax_s \left\{ F_T(t_0 - s) \cdot \frac{\alpha}{(1 - F_T(s))} \right\}$
Очевидно, что $inf(s_r^*) = quant_{F_T}(1 - \alpha)$, то есть достигается на границе, так что оптимум достаточно искать только во втором регионе.
Положим:

$$g(s) = \frac{F_T(t_0 - s)}{(1 - F_T(s))} \tag{4}$$

Таким образом, финальная постановка задачи получилась такой:

$$s^* = argmax_s g(s) \tag{5}$$

При ограничениях:

$$s \leq quant_{F_T}(1-\alpha) \tag{6}$$

Я выкинул α из (4), так как это константа.

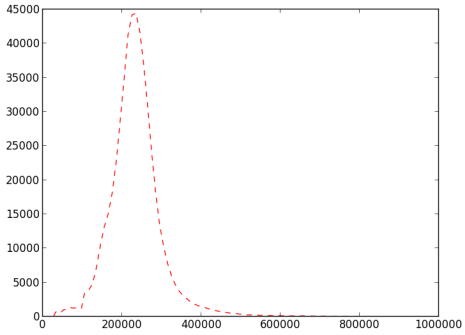
2 Решение

Есть 2 варианта: используем численные методы или делаем предположение о распределении T и аналитически находим условный экстремум g(s).

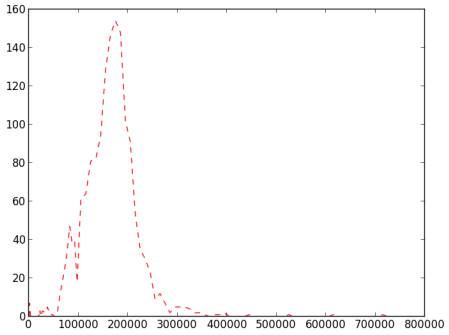
Примеры распределений времен ответов некоторых источников: см Рис. 1 - 5.
Тут сложно сделать предположение, что эти распределения следуют какому-то известному теоретическому распределению, так что пойдем по первому пути.

В нашей задаче вместо "численных методов"мы можем себе позволить перебрать середины всех бакетов гистограммы T (бакетов у нас обычно в районе пары десятков), т.о. приближенное решение, которое должно нас устроить, выглядит так:

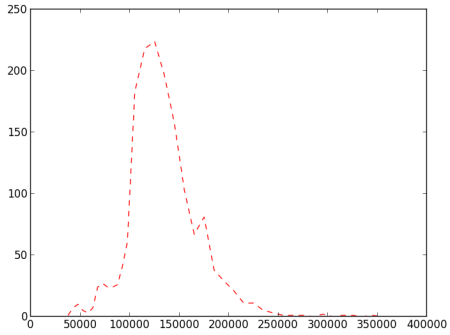
$$\hat{s^*} = max_{i:mid(bucket_i)<quant_{F_T}(1-\alpha)} g(mid(bucket_i)) \tag{7}$$



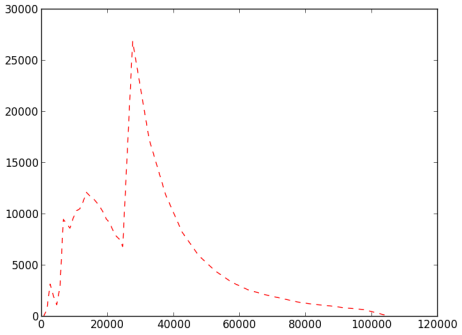
(a) upper-UPPER



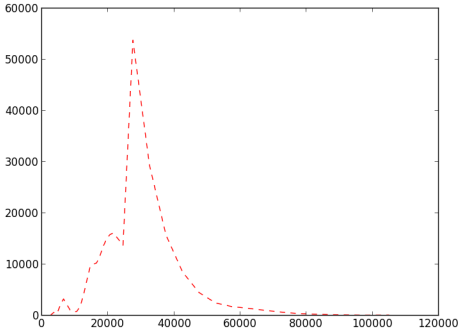
(b) blender5s-WEB



(c) blender5s-UPPER_INT



(d) web4-msp-wiz-parallel-MISSPELL



(e) web4-msp-wiz-parallel-BEGEMOT_GRAPH