Инженерный вестник Дона, №2 (2015)
ivdon.ru/ru/magazine/archive/n2y2015/2900

Методика статистического анализа характеристик входных потоков запросов в системах обработки информации

В.А. Зуев, А.Н. Панфилов, А.Н. Скоба

Южно-Российский государственный политехнический университет (НПИ), Новочеркасск

Аннотация: В статье приводится описание наиболее важных этапов, выполняемых при исследовании входных потоков данных для систем обработки информации.
Ключевые слова: моделирование, запрос, распределение, случайная величина, поток событий, интенсивность потока, статистический анализ, критерий, стационарность, пуассоновский процесс, уровень значимости.

Одной из основных задач определения исходных данных для моделирования процессов обработки информации в распределенных системах обработки информации (СОИ) является нахождение функции *f(t)*, представляющую собой плотность распределения интервалов времени между запросами пользователей СОИ. Эти интервалы составляют случайную последовательность событий и для ее оценки их характеристик необходимо провести процедуру статистического анализа. Исходными данными для статистического анализа являются следующие величины: $x_{i,k}$ - длины временных интервалов между соседними запросами от *k*-го пользователя СОИ, где $i=\overline{1,n_k}$, $k=\overline{1,S_0}$, n_k - объем выборки для *k*-го пользователя, S_0 - общее число пользователей СОИ; $n_{j,k}$ - число запросов от *k*-го пользователя, поступивших в равные промежутки времени $t_{j,k}$, $j=\overline{1,Q_0}$.

На рис.1 показана структура процесса анализа характеристик потока событий. Одной из первоочередных задач статистического анализа потоков событий является проверка независимости и одинаковости распределения случайных величин. Для решения этой задачи используются критерии, основанные на выборочных коэффициентах корреляции и критерии, базирующиеся на спектральной плотности интервалов [1,2].

В соответствии с первым типом критериев, гипотеза о независимости отвергается уровнем значимости α, если $\left|\rho_{j,k}\sqrt{n_k-1}\right| > C_{\alpha/2}$ или $\left|\rho_{j,k}\right| > \frac{C_{\alpha/2}}{\sqrt{n_k-1}}$, где $C_{\alpha/2}$ является верхней (α/2) - квантилью единичного нормального распределения; $\rho_{j,k}$ - коэффициент корреляции *k*-го пользователя СОИ с аргументом запаздывания *j*, который определяется следующим образом:

$$\rho_{j,s} = \frac{C_{j,k}}{\left(C'_{0,j,k}C''_{0,j,k}\right)^{1/2}},$$

где $C_{j,k} = \frac{1}{(n_k-j)}\left(\sum_{i=1}^{n_k-j} x_{i,k}x_{i+j,k}\right) - \bar{x}'_{j,k}\bar{x}''_{j,k}$, $\bar{x}'_{j,k} = \frac{1}{(n_k-j)}\sum_{i=1}^{n_k-j} x_{i,k}$

$$C'_{0,j,k} = \frac{1}{(n_k-j)}\sum_{i=1}^{n_k-j}\left(x_{i,k}-\bar{x}'_{j,k}\right)^2, \quad C''_{0,j,k} = \frac{1}{(n_k-j)}\sum_{i=1}^{n_k-j}\left(x_{i+j,k}-\bar{x}''_{j,k}\right)^2, \quad \bar{x}''_{j,k} = \frac{1}{(n_k-j)}\sum_{i=1}^{n_k-j} x_{i+j,k}.$$

Рис. 1.- Структура процесса анализа характеристик потока событий

Анализ статистических данных с целью установления стационарности потока заявок основан на двух типах методов [1-5]. Первый тип использует стандартные методы наименьшей квадратичной регрессии, а методы второго типа базируются на эффективном теоретическом анализе специальных математических моделей, например, пуассоновского процесса, параметр которого изменяется по некоторому простому закону. Так наиболее эффективным является критерий, предполагающий в качестве нулевой гипотезы пуассоновский процесс, а в качестве конкурирующей - нестационарный пуассоновский процесс с интенсивностью наступления событий вида $\lambda_k(t) = e^{\alpha+\beta t}$, где α и β являются неизвестными параметрами. При этом проверка нулевой гипотезы β =0 для $\lambda_k(t)$ сводится к вычислению выражения:

$$u_k = \frac{\sum_{i=1}^{n_k} t_{i,k} - 0.5 n_k t_{0,k}}{t_{0,k}\sqrt{n_k/12}},$$

где $t_{i,k} = \sum_{r=1}^{i} x_{r,k}$; $t_{0,k} = \sum_{r=1}^{n_k} x_{r,k}$ - период наблюдений.

Нулевая гипотеза принимается, если $u_{\mathbf{k}}$ отличается от нуля менее, чем на 5%. Знак $u_{\mathbf{k}}$ указывает на возрастание или убывание интенсивности.

Одним из стандартных критериев для проверки гипотезы о том, что интервалы $x_{i,k}$ являются наблюденными значениями случайной величины, имеющей показательное распределение с параметром $\lambda_k (k = \overline{1, S_0})$, является дисперсионный критерий, основанный на статистике:

$$d_k = \sum_{i=1}^{n_k} \frac{(x_{i,k} - \bar{x}_k)^2}{\bar{x}_k},$$

где $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n_k} x_{i,k}$.

При нулевой гипотезе распределение величины d^k хорошо аппроксимируется χ^2 распределением с (n^k-1) степенями свободы.

Существует много параметрических семейств функций распределения, которые можно использовать в качестве модели для распределения интервалов времени между событиями процесса восстановления. Наиболее важным из них является распределение Эрланга, плотность распределения которого имеет вид:

$$f(x_k(n_{0,k})) = \frac{\lambda_k(\lambda_k x_k(n_{0,k}))^{(n_{0,k}-1)}\exp(-\lambda_k x_k(n_{0,k}))}{(n_{0,k}-1)!},$$

где $x_k(n_{0,k})$ – время от начала отсчета до генерации $n_{0,k}$ - го по счету запроса k-го пользователя; $n_{0,k}$ –фиксированное целое число, причем $n_{0,k}$ принадлежит отрезку $[0, n_k]$; λ_k – интенсивность формирования запросов k- пользователя.

Для оценки параметров λ_k и $n_{o,k}$ можно воспользоваться критериями χ^2.

В ряде технических приложений, описанных в работе [6], встречаются нестационарные пуассоновские процессы, т.е. процессы, в которых интенсивность наступления событий сама является функцией времени $\lambda(t)$, причем очень часто величина $\lambda(t)$ является реализацией стационарного случайного процесса. Общих методов анализа характеристик таких процессов пока не существует. Единственное общее указание, которое можно сделать относительно анализа потоков событий такого типа, состоит в том, что оценки параметров и проверка гипотез значительно упрощается, если удается обнаружить определенные закономерности процесса (например, спектральную плотность целочисленного процесса).

Среди последовательностей событий, интервалы времени между которыми не являются одинаково распределенными, наибольшее практическое значение имеют так называемые последовательности событий, смещенные случайными воздействиями. Это процессы, в которых события

должны проходить согласно расписанию через определенные интервалы времени, но по различным причинам отклоняются от этих предписанных моментов времени. Наиболее простая модель таких последовательностей получается, если предположить, что согласно расписанию, события должны проходить последовательно через интервал времени a и что задержки являются независимыми и одинаково распределенными случайными величинами. Тогда действительным моментом времени наступления $x_{i,k}$ по расписанию события является: $t_{i,k} = a_0 + ka + b_k$. Здесь b_k является реализацией некоторой случайной величины B, с функцией распределения $F_B(x_{i,k})$ и дисперсией σ_B^2. В работе [6] приведены основные соотношения для статистической оценки корреляции интервалов времени между событиями такого типа.

Для сравнения интенсивностей потоков запросов от каждого пользователя СОИ можно использовать критерии, основанные на отношении функции максимального правдоподобия и индексе дисперсии [1]. Нулевая гипотеза состоит в равенстве $\lambda = \lambda_k$, а конкурирующая гипотеза предполагает различную интенсивность для каждого из k пользователей распределенной СОИ. При нулевой гипотезе случайная величина $H = \left(\sum_{k=1}^{S_o} n_k \cdot \ln(n_k / t_{0,k}) - \sum_{k=1}^{S_o} n_k \cdot \ln(\sum_{k=1}^{S_o} n_k / \sum_{k=1}^{S_o} t_{o,k}) \right)$ имеет χ^2 распределение с $(S_0 - 1)$ степенями свободы. При малом уровне значимости случайной величины H, не позволяющем сделать окончательных выводов о справедливости нулевой гипотезы, равенство интенсивностей потоков проверяется по критерию индекса дисперсии [1,2].

Основные этапы предложенной методики статистического анализа входных потоков были реализованы в среде MatLab [7,8] и использованы для оценивания потока запросов пользователей информационных систем

организационного управления. Экспериментальные данные подтверждают гипотезу о стационарности, независимости и экспоненциальной плотности распределения времени между запросами.

Литература

1. Кокс Д., Льюис П. Статистический анализ последовательности событий. М.: Мир, 1969. 312с.

2. Бендат Дж., Пирсол А. Прикладной анализ случайных данных. М.: Мир, 1989. 540с.

3. Андерсон Т. Статистический анализ временных рядов. М.: Мир, 1976. 755 с.

4. Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, 820 p.

5. Большаков И. А., Ракошиц В. С. Прикладная теория случайных потоков. М.: Сов.радио, 1978. 248 с.

6. Оран Э., Борис Дж. Статистическое моделирование реагирующих потоков. М.:Мир, 1990. 390с.

7. Martinez, W.L. and A.R. Martinez, 2002. Computational Statistics Handbook with MATLAB. London: CHAPMAN & HALL/CRC, 763 p.

8. Дьяконов В. MATLAB: учебный курс. СПб: Питер, 2001. 560с.

9. Зырянов В.В. Методы оценки адекватности результатов моделирования // Инженерный вестник Дона, 2013, №2 URL:ivdon.ru/ru/magazine/archive/n2y2013/1707/.

10. Якоб Д.А. Разработка алгоритма нахождения входного потока заявок в имитационной модели контрольно-пропускной системы на основе статистических данных //Инженерный вестник Дона, 2014, №3 URL:ivdon.ru/ru/magazine/archive/n3y2014/2480/.

References

1. Koks D., L'yuis P. Statisticheskiy analiz posledovatel'nosti sobytiy. M.: Mir, 1969. 312 p.

2. Bendat Dzh., Pirsol A. Prikladnoy analiz sluchaynykh dannykh [Random Data. Aanalysis and Measurement Procedure]. M.: Mir, 1989. 540 p.

3. Anderson T. Statisticheskiy analiz vremennykh ryadov [Statistical analysis of temporary ranks]. M.: Mir, 1976. 755 p.

4. Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, 820 p.

5. Bol'shakov I. A., Rakoshits V. S. Prikladnaya teoriya sluchaynykh potokov. M.: Sov. radio, 1978. 248 p.

6. Oran E., Boris Dzh. Statisticheskoe modelirovanie reagiruyushchik hpotokov. M.: Mir, 1990. 390 p.

7. Martinez, W.L. and A.R. Martinez, 2002. Computational Statistics Handbook with MATLAB. London: CHAPMAN & HALL/CRC, 763 p.

8. D'yakonov V. MATLAB: uchebnyy kurs. SPb: Piter, 2001. 560 p.

9. Zyryanov V.V. Inženernyj vestnik Dona (Rus), 2013, №2 URL:ivdon.ru/ru/magazine/archive/n2y2013/1707/.

10. Yakob D.A. Inženernyj vestnik Dona (Rus), 2014, №3 URL:ivdon.ru/ru/magazine/archive/n3y2014/2480/.