Информатика и вычислительная техника

УДК 004.89 **DOI: 10.14529/ctcr180301**

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ ДЛЯ ПРЕДСКАЗАНИЯ РАСХОДА ПАМЯТИ В ВЫСОКОНАГРУЖЕННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ

А.В. Тузов

Челябинский государственный университет, г. Челябинск, Россия

Рассматривается актуальная на данный момент проблема планирования задач в высоконагруженных информационных системах. Целью данной работы является проверка гипотезы о том, что загруженность высокопроизводительных информационных систем зависит от внешних параметров среды, в которой они функционируют. Для проверки была собрана и запущена система, на которой находятся корпоративный сайт компании, система мониторинга и приложение для социальной сети vk.com. В качестве внешних параметров были выбраны как природные явления, так и статистические данные посещения популярных сайтов, а также курсы валют и акций. На наш взгляд, эти параметры в той или иной степени могут оказывать влияние на загруженность информационной системы. Данные собирались на протяжении месяца работы системы каждые десять минут. При каждом сборе информации для каждого работающего процесса в системе запоминалось количество расходуемой им памяти. Для идентификации модели был выбран метод линейной регрессии как наиболее простой и часто используемый вариант проверки неявных зависимостей между данными. Все собранные параметры были отфильтрованы – проверены на наличие кросскорреляции и нормализованы. Используя построенную модель, мы предсказали значение расходуемой памяти процессами. Для каждого предсказанного значения было посчитано среднеквадратичное отклонение. Анализ результатов показал, что построенная модель имеет ряд проблем. В качестве рекомендаций по улучшению результатов указано использование другого метода построения модели, а также улучшение качества и количество собираемых данных. Дальнейшие планы включают в себя исследование возможности предсказания процессорного времени высоконагруженной информационной системы, используя внешние параметры.

Ключевые слова: машинное обучение, линейная регрессия, процессы операционной системы, оперативная память.

Введение

В последние годы резко выросло число прикладных задач, обеспечивающих сервисами тысячи и сотни тысяч пользователей. Для решения задачи получения приемлемого времени отклика на запросы пользователей используется специальный класс вычислительных систем, которые получили название высоконагруженные (Highload application). Для подобных систем характерны следующие свойства: большое количество одновременных пользователей, большой объем обрабатываемой информации. Однако главным критерием является масштабируемость, т. е. доступность для любого теоритически достижимого числа клиентов при сохранении приемлемого времени отклика. На данный момент сложности, возникающие в системах, связанные с расходом процессорного времени, расходом оперативной памяти [1] и планировании стека задач, как правило, решаются добавлением в систему вычислительных мощностей либо попыткой распараллелить вычисления, что не всегда просто и опять же требует дополнительных мощностей. В некоторых работах изучается возможность изменения архитектуры операционной системы, в частности, отказ от ядра при операциях с данными ввода-вывода в контексте операционной системы общего назначения, для повышения производительности, сохраняя при этом традиционную модель безопасности [2]. Однако экстенсивный путь развития является не единственным. За счет грамотного

планирования и распределения ресурсов высоконагруженных систем (ВС) можно обойтись меньшими мощностями [3]. Ряд работ исследует возможность предсказания краткосрочной нагрузки для эффективной работы энергосистемы с помощью внешних параметров (температура воздуха, осадки, скорость ветра и прочее) [4–6]. В данной работе сделана попытка проверить гипотезу о том, что загруженность информационных систем зависит от внешних параметров окружающей среды, в которой они функционируют. Проверка гипотезы будет осуществляться с помощью алгоритмов машинного обучения, используемых для выявления неявных зависимостей между данными.

1. Теоретическая часть

Традиционно под процессом понимается системная или прикладная программа, находящаяся на стадии выполнения, с которой связаны определенное состояние памяти, значения общих регистров процессора, состояния открытых файлов, текущий каталог и прочее [1]. Задачей операционной системы является управление процессами и ресурсами компьютера или, точнее организация рационального использования ресурсов в интересах наиболее эффективного выполнения процессов [7]. Для решения подобной задачи операционная система должна знать об имеющихся у нее ресурсах и о том, какой процесс обладает и работает с какими ресурсами. Основной подход к хранению такой информации заключается в создании и поддержке таблиц, содержащих данную информацию.

С каждым процессом связывается его адресное пространство – набор адресов в памяти, в который процесс может писать и читать данные. Ресурсом, расход которого мы будем предсказывать в данной работе, является память. Память – это совокупность регистров, кэша, операционной и дисковой памяти [1]. Для работы процесс или часть его должна находиться в операционной памяти. Однако, как правило, объема памяти не достаточно для запуска в операционной памяти всех процессов системы, потому существуют два способа избежания перегрузки памяти – это свопинг и виртуальная память. С учетом того, что в высоконагруженных системах количество процессов может различаться на порядок, нужно грамотно выбрать алгоритм планирования процессов. На сегодняшний момент наиболее популярными являются: First-Come, First-Served (FCFS), Карусель (Round Robin – RR), Shortest-Job-First (SJF), Гарантированное планирование, Приоритетное планирование, Многоуровневые очереди, Многоуровневые очереди с обратной связью [1, 8, 9]. Материалов по использованию машинного обучения для планирования стека процессов крайне мало [10, 11].

Подавляющее количество исследований, затрагивающих использование машинного обучения для предсказания расхода ресурсов в компьютерных системах, касается энергетической сферы. В последние годы набирают популярность исследования, которые используют внешние признаки окружающей среды для предсказания расходов в электрических системах [12, 13]. Принципиальное отличие этой работы в том, что для высоконагруженных информационных систем таких исследований не проводилось. Для построения модели был выбран метод линейной регрессии, так как он является наиболее простым, часто используемым и наиболее изученным.

Линейная модель [14–16] имеет вид:

$$y = Xb + \varepsilon,$$

где $y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$ – вектор значений зависимой переменной y, где y – значение расходуемой процессом оперативной памяти в килобайтах;

$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$ – матрица внешних параметров;

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$ – вектор случайных ошибок.

Так как количество внешних параметров, которое мы использовали, равно 18, то и $k = 18$. Сбор данных производился каждый десять минут в течение месяца, поэтому значение $n =$ 285 376. В нашем случае вектором случайных ошибок можно пренебречь.

Полный список параметров представлен в экспериментальной части, здесь мы опишем причины выбора этих параметров. По нашему мнению, эти параметры оказывают влияние на информационную систему.

- Температура воздуха в городе, в котором расположена система, а также температура воздуха в крупных городах влияют на увеличение загруженности системы. Чем ниже температура, тем выше нагрузка на ВС.
- Большое количество новостей и постов на популярных сайтах вызывают интерес пользователей, они больше времени проводят на информационных системах, изменяя рейтинг постов, где это возможно. Соответственно, чем больше публикуемых новостей, постов и их рейтингов, тем больше нагрузка на ВС.
- Пропускная способность каналов связи оказывает влияние на нагрузку, оказываемую на информационные высоконагруженные системы. Чем выше входящая и исходящая скорости соединения, тем больше посещений. Данную тенденцию можно было наблюдать в России и Китае на протяжении 2000–2010 годов, период, в течение которого количество и качество каналов связи росло. Это незамедлительно сказалось на количестве запросов к ВС.
- Оказываясь в большой «пробке», пользователи пытаются занять себя чем-либо. В связи с распространением, небольшой стоимостью и возросшим качеством мобильного интернета посещение информационных систем стало основным способом проведения времени в «пробке». Соответственно, чем больше уровень «пробок» и чем их больше, тем выше нагрузка на ВС.
- Чем не стабильнее стоимость акций компаний, курсов валют, нефти и драгоценных металлов, которыми пользуются миллиарды людей, тем больше людей находятся в интернете, покупая, продавая или отслеживая курсы. Это наглядно было продемонстрированно в разгар кризиса 2015 года в России и, соответственно, роста доллара и евро, а также в 2017 года в связи с ростом биткоина.

С учетом того, что мы берем информацию из сторонних источников, то надо убедиться в ее достоверности. Rss лента сайтов, которые мы берем для подсчета количества новостей, не вызывает сомнения, так как крупные сайты либо добавляют rss в автоматическом режиме, либо используют rss для добавления новостей. Информация о курсах иностранных валют по отношению к рублю, опубликованная на официальном сайте Банка России в сети Интернет, является официальной информацией Банка России и не требует дополнительного письменного подтверждения от Банка России. Биржевые данные о торгах акциями, курсом нефти и золота, размещенные на сайте, предоставляют информацию с Московской Биржи. Рейтинги постов, значение температуры и уровень пробок рассчитываются с помощью внутренних средств сайтов, информация о которых является конфиденциальной и отсутствует в открытых источниках информации.

Для каждого набора параметров будет считаться среднеквадратичное отклонение, которое рассчитывается по формуле

$$A = \sqrt{\frac{\sum_{i=0}^{n}(y1_i - y2_i)^2}{n}}, \quad (1)$$

где, $y1_i$ – предсказанное значение оперативной памяти, $y2_i$ – реальное значение оперативной памяти, n – размер собранной выборки.

2. Экспериментальная часть

Имеется сервер с операционной системой linux ubuntu 14.04.5 LTS. Процессор Intel(R) Xeon(R) CPU X5650 2.67 GHz. 16 гигабайт оперативной памяти, 105 гигабайт жесткий диск. На сервере работают сервисы apache и mysql для поддержки корпоративного сайта компании и приложения для vk.com и zabbix, которые следят за ресурсами на сервере.

Были собраны такие внешне параметры:

- cbr.ru – курсы доллара, евро и юаня;
- finanz.ru – курс акций Яндекса, Сбербанка, Газпрома, Yahoo, Microsoft, Google;
- finanz.ru – стоимость марки brent, курс золота и серебра;

• gismeteo.ru – значение температуры в Челябинске, Москве, Екатеринбурге и Санкт-Петербурге;

• www.lenta.ru – количество новостей на текущий момент времени;

• www.pikabu.ru – рейтинг самого большого поста;

• 2ip.ru – самая большая входящая и исходящая скорость, название провайдера и самый маленький пинг;

• habrahabr.ru – количество новых публикаций;

• autochel.ru – уровень пробок в Челябинске;

• reestr.rublacklist.net – количество запрещенных сайтов на портале Роскомнадзора;

• st.kp.yandex.net – количество новостей.

Количество новостей на сайтах:

• st.kp.yandex.net;

• fakty.ua;

• feeds.bbci.co.uk;

• un.org;

• securitylab.ru;

• sports.ru;

• 3dnews.ru;

• osp.ru.

Данные собирались каждые 10 мин в течение месяца. Общее количество строк в файле составило 285376. В собранном файле имена были заменены идентификаторами. Память процессов с одинаковыми идентификаторами, собранными за проход, была просуммирована. Параметры, состоящие из текста, были отброшены, и итоговая строка имела вид, где последнее значение в строке занимало значение памяти в килобайтах, расходуемое процессом в данный период времени.

Проанализировав получившуюся выборку, были удалены параметры, которые не менялись за время сборки это: количество новостей с сайта lenta.ru, kinopoisk.ru, fakty.ua, un.org, securitylab.ru, 3dnews.ru, osp.ru. Все оставшиеся параметры надо было проверить на наличие корреляции относительно друг друга с использованием критерия Пирсона. В результате проверки были получены значения, превышающие средний уровень корреляции.

Все остальные значения не превышали средней корреляции, т. е. 0,7, и не оказывают сильного влияния друг на друга, а значит их совместное использование для построения модели корректно. Из пар параметров, представленных в таблице, были оставлены такие параметры, как юань, серебро и стоимость акции ПАО «Сбербанка».

Таблица

Первый параметр	Второй параметр	Значение корреляции
Доллар	Евро	0,77
Доллар	Юань	0,99
Юань	Евро	0,82
Золото	Серебро	0,76
Акции ПАО «Газпром»	Акции ПАО «Сбербанка»	0,97

Получившийся файл мы разделили на два. Первые файл из 190 250 строк – обучающая выборка, второй из 95 124 – проверочная. Для построения математической модели был выбран метод линейной регрессии. Из-за того, что некоторые значения параметров меньше единицы, а другие значения превышают тысячу, была произведена нормализация данных через логарифмическое преобразование. Наилучший результат показала модель с параметрами, при которых среднеквадратичное отклонение (1) получилось 2,58 килобайт. Для высоконагруженной информационной системы данное отклонение не является оптимальным, так как в условиях нехватки ресурсов, в нашем случае оперативной памяти, это может сильно увеличить время отклика.

Анализ показывает, что построенная модель имеет ряд проблем. Возможно, сбор большего числа признаков, а также использование других методов построения моделей уменьшит среднеквадратичное отклонение. Данное исследование включало такие параметры, как рейтинги и новости сайтов, которые напрямую зависят от влияния человека и которые не имеют естественного происхождения. Возможно, для узкоспециализированных систем, где внешние естественные признаки меняются часто и сильнее влияют на работу системы, среднеквадратичное отклонение (1) математической модели будет меньше. Качество собираемых данных также стоит улучшить, что значительно скажется на качестве построенной модели.

Заключение

За последние годы возросло число сервисов, которыми пользуются сотни тысяч людей, а соответственно появилось много ВС. Однако в таких системах имеется ряд проблем, связанных с расходом ресурсов (процессорного времени, оперативной памяти и прочее). В данной работе была проверена гипотеза о том, что загруженность информационных систем зависит от внешних параметров окружающей среды, в которой они функционируют. Для проверки гипотезы были использованы алгоритмы машинного обучения. По результатам исследования показана принципиальная зависимость параметров ВС от внешних параметров, что позволяет построить эффективный алгоритм планирования сервисных процессов ВС. На основании внешних параметров с использованием метода линейной регрессии была построена математическая модель для предсказания расходов операционной памяти. Для модели была посчитано среднеквадратичное отклонение (1), которое составило 2,59 килобайт. Анализ показывает, что нужно улучшать качество и количество собираемых данных. Также из-за большого отклонения стоит использовать более сложные методы построения модели. Для будущей работы следует увеличить количество внешних параметров, при этом стоит собирать естественные параметры – температуру воздуха, атмосферное давление и прочие, отказавшись от тех, на которые влияет человек. Также следует попробовать построить математическую модель для предсказания процессорного времени процесса с использованием внешних параметров окружающей среды, в которой функционирует ВС.

Литература/References

1. Окороков В.А. Операционные системы: курс лекций. Челябинск: Изд-во Челябинского гос. ун-та, 2011. 288 с. [Okorokov V.A. *Operacionnye sistemy: kurs lektsiy* [Operating Systems: Course of Lectures]. Cheliabinsk, ChSU Publ., 2011. 288 p.]

2. Peter S., Jialin Li, Zhang I., Dan R. K. Ports, Woos D., Krishnamurthy A., Anderson T., Roscoe T. Arrakis: The Operating System Is the Control Plane. *ACM Transactions on Computer Systems*, 2015, vol. 33, no. 4, article 11.

3. Yang R., Ouyang X., Chen Y., Townend P., Xu J. Intelligent Resource Scheduling at Scale: a Machine Learning Perspective. *IEEE International Symposium on Service Oriented System Engineering*, 2018, pp. 132–141. DOI: 10.1109/SOSE.2018.00025

4. Zheng H., Yuan J., Chen L. Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies*, 2017, vol 10, no. 8. Available at: http://www.mdpi.com/1996-1073/10/8/1168/htm (accessed 1 August 2017). DOI: 10.3390/en10081168

5. Divina F., Gilson A., Goméz-Vela F., García Torres M., & Torres J.F. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. *Energies*, 2018, vol. 11, no. 4. Available at: http://www.mdpi.com/1996-1073/11/4/949/htm (accessed 9 April 2018). DOI: 10.3390/en11040949

6. Dahua Gan, Yi Wang, Ning Zhang, Wenjun Zhu. Enhancing Short-Term Probabilistic Residential Load Forecasting with Quantile Long-Short-Term Memory. *The Journal of Engineering*, 2017, vol. 2017, iss. 14, pp. 2622–2627. DOI: 10.1049/joe.2017.0833

7. Назаров С.В., Широков А.И. Современные операционные системы. М., 2012. 367 с. [Nazarov S.V., Shirokov A.,I. *Sovremennie operatsionnye systemi* [Modern Operating Systems]. Moscow, 2013. 367 p.]

8. *Fair scheduler (2018).* Available at: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/FairScheduler.html (accessed 16 April 2018).

9. *Capacity scheduler(2018)*. Available at: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/CapacityScheduler.html (accessed 16 April 2018).

10. Wang F., Gao X., Chen G. Lowering the Volatility: a Practical Cache Allocation Prediction and Stability-Oriented Co-Runner Scheduling Algorithms. *The Journal of Supercomputing*, 2017, vol. 72, no 3, pp. 1126–1151. DOI: 10.1007/s11227-016-1645-7

11. Evans R., Gao J. DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. *DeepMind Blog* (2016), vol. 20. Available at: https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/ (accessed 20 July 2016).

12. Bećirović E., Ćosović M. Machine Learning Techniques for Short-term load Forecasting. *Environment Friendly Energies and Applications (EFEA), 4th International Symposium*, 2016, pp. 1–4.

13. Kim T., Lee D., Choi J., Spurlock A., Sim A., Todd A., Wu K. Extracting Baseline Electricity Usage with Gradient Tree Boosting. *Smart City/SocialCom/SustainCom (SmartCity), IEEE International Conference*, 2015, pp. 734–741. DOI: 10.1109/SmartCity.2015.156

14. Freedman D.A. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009. 456 p. DOI: 10.1017/CBO9780511815867

15. Neter J., Kutner M.H., Nachtsheim C.J., & Wasserman W. *Applied Linear Statistical Models*. Chicago: Irwin, 1996, vol. 4. 318 p.

16. Rao C.R., Toutenburg H. Linear Models. *Linear Models: Least Squares and Alternatives*. Springer, 1995, pp. 3–18. DOI: 10.1007/978-1-4899-0024-1, DOI: 10.1007/978-1-4899-0024-1_2

Тузов Артем Викторович, аспирант, Челябинский государственный университет, г. Челябинск; amirel92@mail.ru.

Поступила в редакцию 23 апреля 2018 г.

DOI: 10.14529/ctcr180301

INVESTIGATION OF THE POSSIBILITY OF USING LINEAR REGRESSION FOR PREDICTING MEMORY CONSUMPTION IN A HIGHLOAD INFORMATION SYSTEM

***A.V. Tuzov**, amirel92@mail.ru*

Chelyabinsk State University, Chelyabinsk, Russian Federation

The article considers the actual problem of planning tasks in highloaded information systems at the moment. The purpose of this paper is to test the hypothesis that the congestion of high-performance information systems depends on the external parameters of the environment in which they operate. For verification, the system on which the corporate website of the company, the monitoring system and the application for the social network vk.com were collected and launched. As external parameters were chosen as natural phenomena, as well as statistical data of visiting popular sites, as well as exchange rates and shares. In our opinion, these parameters may to some extent influence the workload of the information system. The data was collected during the month of the system operation every ten minutes. At each collection of information for each running process in the system, the amount of memory it consumes is remembered. To identify the model, the linear regression method was chosen, as the most simple and often used option for verifying implicit dependencies between data. All the collected parameters were filtered out – checked for cross-matching and normalized. Using the constructed model, we predicted the value of memory consumed by processes. For each predicted value, the root-mean-square deviation was calculated.

Analysis of the results showed that the model constructed has a number of problems. As recommendations for improving the results, the use of another method to build a model is indicated, as well as improvement of the quality and quantity of data collected. Further plans include exploring the possibility of predicting the CPU time of a highload information system using external parameters.

Keywords: machine learning, linear regression, operating system process, random access memory.

Received 23 April 2018

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Тузов, А.В. Исследование возможности использования линейной регрессии для предсказания расхода памяти в высоконагруженной информационной системе / А.В. Тузов // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2018. – Т. 18, № 3. – С. 5–11. DOI: 10.14529/ctcr180301

FOR CITATION

Tuzov A.V. Investigation of the Possibility of Using Linear Regression for Predicting Memory Consumption in a Highload Information System. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2018, vol. 18, no. 3, pp. 5–11. (in Russ.) DOI: 10.14529/ctcr180301