Использование регрессии для предсказания расходов памяти в высокопроизводительных информационных системах

А.В. Тузов
Институт информационных технологий
Челябинский государственный университет
Челябинск, Россия
e-mail: amirel92@mail.ru

Аннотация[1]

В данной работе мы попытаемся проверить гипотезу о том, возможно ли эффективно использовать машинное обучение для предсказания расхода ресурсов процессами в информационной высоконагруженной системе. Для идентификации модели используется метод линейной регрессии.

1. 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. Главным критерием данных систем является масштабируемость, т.е. доступность для любого теоретически 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 при этом, сохраняя 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. Ряд работ исследует возможность предсказания краткосрочной нагрузки для эффективной работы энергосистемы с помощью внешних параметров (температура воздуха, осадки, скорость ветра и прочее) [3]. 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, используемые для выявления неявных зависимостей между данными.

2. 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 [4]. 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 поддержки таблиц, содержащих данную информацией.

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]. Материалы по использованию машинного обучения для планирования стека процессов отсутствуют.

Труды Шестой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-31 мая, Уфа-Ставрополь, Россия, 2018

Линейная модель[5] имеет вид:

$$y = Xb + \varepsilon \quad (1)$$

где y – вектор значение расходуемой процессом оперативной памяти в килобайтах, X – матрица внешних параметров и ε – вектор случайных ошибок

Был выбран список параметров который по нашему мнению, оказывают влияние на информационную систему:

- Значение температуры воздуха в городе, в котором расположена система, а также температура воздуха в крупных городах, влияет на увеличение загруженность системы. Чем ниже температура, тем выше нагрузка на ВС;
- Большое количество новостей и постов на популярных сайтов вызывают интерес пользователей, они больше времени проводят время на информационных системах, изменяя рейтинг постов, где это возможно. Соответственно, чем больше публикуемых новостей, постов и рейтингов постов, тем больше нагрузка на ВС;
- Пропускная способность каналов связи оказывает влияние на нагрузку, оказываемую на информационные высоконагруженные системы. Чем выше, входящая и исходящая скорость 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. Что 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 проведением 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. Что наглядно было продемонстрированно в разгар кризиса 2015 года в России и, соответственно, роста доллара и евро. А так же в 2017 года в связи с ростом биткоина;

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

$$A = \sqrt{\frac{\sum_{i=0} (\kappa_i - в_i)^2}{n}} \quad (2)$$

где k_i – предсказанное значение оперативной памяти, b_i – реальное значение оперативной памяти и n – размер собранной выборки.

3. Экспериментальная часть

Для анализа были собраны такие параметры как:

1. cbr.ru – курсы доллара, евро и юаня;
2. finanz.ru – курс акций Яндекса, Сбербанка, Газпрома, Yahoo, Microsoft, Google;
3. finanz.ru – стоимость марки brent, курс золота и серебра;
4. gismeteo.ru – значение температуры в Челябинске, Москве, Екатеринбурге и Санкт-Петербурге;
5. www.lenta.ru – количество новостей на текущий момент времени;
6. www.pikabu.ru – рейтинг самого большого поста;
7. 2ip.ru – самая большая входящая и исходящая скорость, название провайдера и самый маленький пинг;
8. habrahabr.ru – количество новых публикаций;
9. autochel.ru – уровень пробок в Челябинске;
10. reestr.rublacklist.net – количество запрещенных сайтов на портале Роскомнадзора
11. 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 285 376. 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 сроке 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, которые были отсеяны Получившийся файл мы разделили на два обучающая и 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 (2) получилось 2,58 килобайт. Для высоконагруженной информационной системы данное отклонение не является оптимальным, т. к. 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 так же 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 (2) математической модели будет меньше. Качество собираемых данных так же стоит улучшить, что значительно скажется на качестве построенной модели.

4. 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 (2), которое составило 2,59 килобайт. Анализ показывает, что нужно улучшать качество и количество собираемых данных. Так же 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 – температура воздуха, атмосферное давление и прочие, отказавшись от тех, на которые влияет человек. Так же следует попробовать построить математическую модель для предсказания процессорного времени процесса с использованием внешних параметров окружающей среды, в которой функционирует ВС.

Список используемых источников

1. Операционные системы: Курс лекций/ В.А. Окороков. — Челябинск: Издательство Челябинского государственного университета, 2011. — 287с.
2. Peter S. et al. Arrakis: The operating system is the control plane // ACM Transactions on Computer Systems (TOCS). 2016. Vol. 33. №. 4. – P. 11.
3. Huiting Zheng, Jiabin Yuan, Long Chen. Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation // Energies. 2017. Vol. 10. №. 8. P. 1168.
4. Современные операционные системы/ С.В. Назаров, А.И. Широков — Москва: Бином Лаборатория знаний, 2012
5. Statistical Models: Theory and Practice / David A. Freedman. Cambridge: Cambridge University Press, 2009
6. Bećirović E. Machine learning techniques for short-term load forecasting/ Elvisa Bećirović, Marijana Ćosović // Environment Friendly Energies and Applications (EFEA), 2016 4th International Symposium on. – IEEE, 2016. – P. 1-4.
7. Kim T. Extracting Baseline Electricity Usage Using Gradient Tree Boosting / Taehoon Kim, Dongeun Lee, Jaesik Choi, Anna Spurlock, Alex Sim, Annika Todd, Kesheng Wu //Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on. – IEEE, 2015. – P. 734-741.