Декабрь, № 8 **Физико-математические науки** **2013**

УДК 004.032.26

ОЛЕГ ВАЛЕРЬЕВИЧ ГУСЕВ
преподаватель кафедры прикладной математики и кибернетики математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
eleset@gmail.com

АРТЕМ ВЛАДИМИРОВИЧ ЖУКОВ
кандидат технических наук, старший преподаватель кафедры прикладной математики и кибернетики математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
artem.v.zhukov@gmail.com

ИРИНА ВАЛЕРЬЕВНА ПЕШКОВА
кандидат физико-математических наук, доцент, и. о. заведующего кафедрой прикладной математики и кибернетики математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
iaminova@psu.karelia.ru

СПОСОБ ИДЕНТИФИКАЦИИ ПЕРЕГРУЗКИ WEB-СЕРВЕРА ПРИ ПОМОЩИ НЕЙРОННОЙ СЕТИ*

Рассматривается проблема диагностирования и предотвращения перегрузки при работе web-серверов. Установлено, что использование простых моделей не позволяет добиться точной идентификации перегрузки, а использование сложных – приемлемого уровня производительности алгоритма идентификации. К тому же при использовании реального web-сервера могут проявляться особенности, присущие конкретной аппаратно-программной платформе и оказывающие непосредственное влияние на ее способность противостоять перегрузкам. Предложен подход, основанный на идентификации состояния перегрузки сервера с использованием нейронной сети с архитектурой трехслойного персептрона прямого распространения. Обозначен круг исходных данных, необходимый для формирования набора входных данных нейронной сети. В работе предлагается алгоритм сбора входных данных, обработки полученных результатов и их применения для прогнозирования возникновения состояния перегрузки на web-сервере. Предложенный подход апробирован путем анализа данных работы web-сервера, полученных в ходе нагрузочного тестирования. Полученные результаты показали, что предложенный подход позволяет более точно предсказать состояние перегрузки, что дает возможность организовывать качественное управление запросами web-сервера с целью предотвращения его перегрузки и обеспечения стабильности работы.

Ключевые слова: перегрузка, нейронные сети, управление запросами, web-сервер, нагрузочное тестирование, запрос

К преимуществам реализации информационных систем в виде web-сервисов можно отнести минимальные требования к пользовательской части, возможность эффективно решать сложные ресурсоемкие задачи благодаря значительным аппаратным ресурсам сервера, отработанную технологию разработки подобных систем и т. д. Однако использование web-сервисов связано с необходимостью преодоления ряда сложностей, в частности с обеспечением устойчивой работы в режимах предельных нагрузок, то есть в ситуациях, когда интенсивность поступающих запросов к web-серверу (сервера) превышает имеющиеся возможности по их обработке [3].

Под состоянием перегрузки будем понимать состояние сервера, при котором время обработки запросов превышает некоторую установленную величину L и запрос не получает информационную услугу (получает отказ) вследствие недостатка аппаратных ресурсов.

Несмотря на то что состояние перегрузки может достаточно успешно диагностироваться со стороны пользователей системы, прогнозирование возникновения состояния перегрузки на стороне сервера не всегда является тривиальной задачей. Прогноз возникновения состояния перегрузки сервера строится в момент принятия решения о дальнейшей обработке запроса, который может вызвать перегрузку, на основе имеющихся сведений о количестве обрабатываемых запросов и состоянии сервера.

Как правило, для решения такой задачи используется подход, основанный на количественном определении числа запросов, вызывающих перегрузку, и вычислении функции ресурсоемкости [2]. На практике такой подход не всегда применим ввиду того, что современный web-сер-

вер, как правило, является приложением, выполняемым в рамках контекста операционной системы, самостоятельно распределяющей ресурсы сервера среди множества различных задач, а выделить из общего количества используемых сервером ресурсов лишь те, которые применяются для обработки запроса, практически невозможно. В связи с этим актуальной становится задача идентификации состояния перегрузки с учетом состояния сервера и его текущей загруженности. Для решения данной проблемы предлагается использование аппарата нейронных сетей [4].

Архитектура нейронной сети представляет собой трехслойный персептрон с одним скрытым слоем, входные данные которого отражают состояние сервера в конкретный момент времени, а выходные данные – время обработки запроса или его превышение над заданной величиной L (служат для идентификации состояния перегрузки).

Способность скрытых нейронов выделять статистические зависимости порядка особенно существенна, когда размер входного слоя достаточно велик, что актуально для решения нашей задачи.

В качестве входных данных использовались два вида набора значений:

- Данные, характеризующие загруженность программной части web-сервера как исполняемой программной задачи, а именно данные о количестве запросов, которые находятся в обработке в разрезе каждого сетевого сервиса.
- Данные, характеризующие загруженность сервера в целом как программно-аппаратного комплекса, распределяющего аппаратные ресурсы между всеми задачами, а именно предоставляемые операционной системой данные, характеризующие основные показатели работы сервера (счетчики производительности).

Выходной сигнал сигнализировал о наличии или отсутствии перегрузки в условиях заданного состояния сервера.

Для практического применения нейронной сети к задаче идентификации перегрузки необходимо решить проблему ее первоначального обучения и актуализации параметров модели в процессе использования.

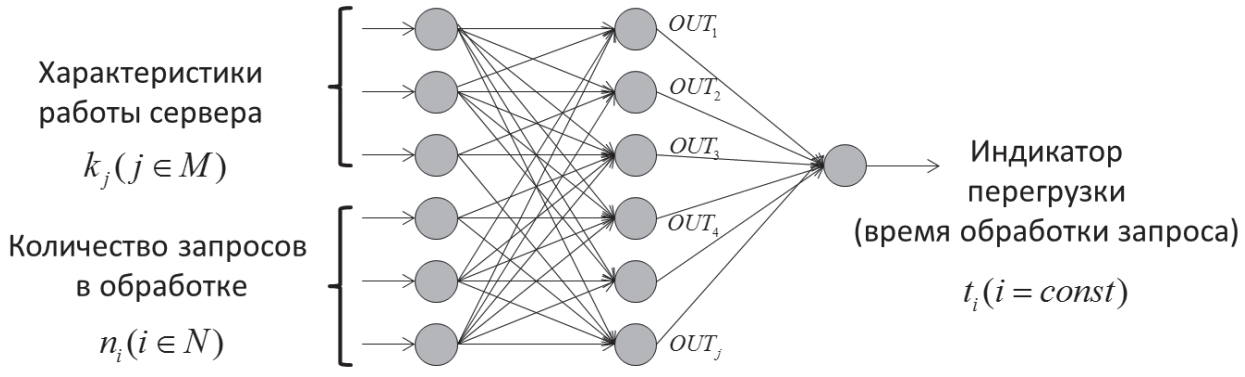
В качестве входных данных для расчета используются показатели, характеризующие работу сервера и предоставляемые операционной системой, а именно количество запросов в обработке на сервере к каждому из функционирующих сетевых сервисов и индикативные величины состояния ресурсов сервера (счетчики производительности), позволяющие охарактеризовать уровень его загруженности (см. рисунок).

В качестве функции активации нейронов промежуточного и выходного слоя использовалась логистическая функция вида

$$OUT_d = \frac{1}{1+\exp\left(-\alpha\left(\sum_{j\in M} w_{1j}k_j + \sum_{i\in N} w_{2i}n_i\right)\right)},$$

где OUT_d – выходной сигнал искусственного нейрона; α – параметр наклона сигмоидальной функции активации; M – множество величин, характеризующих загруженность аппаратных ресурсов сервера ($j \in M$), существенных для оценки времени выполнения запросов; $\overline{M}$ – множество величин, характеризующих загруженность аппаратных ресурсов сервера ($M \subset \overline{M}$); N – множество сетевых сервисов на сервере ($j \in N$); k_j – индикативная величина загруженности аппаратного ресурса сервера ($j \in M$); n_i – количество запросов в обработке к сетевому сервису ($j \in N$); w_{1j}, w_{2i} – синаптические веса.

Выбор индикативных величин должен быть таким, чтобы их значения оказывались в статистической связи с индикатором перегрузки, то есть временем обработки запроса. Одним из характеризующих параметров могло бы стать использование коэффициента линейной корреляции для установления такой связи. В то же время набор индикаторов в максимальной степени должен представлять собой набор базисных компонент. Поскольку набор влияющих факторов в общем случае различен для разных



Нейронная сеть прямого распространения для решения задачи идентификации перегрузки web-сервера

аппаратно-программных платформ, правильный выбор индикативных величин во многом может определить эффективность идентификации перегрузки для конкретного сервера.

Поскольку предельное время обработки запроса к каждому сетевому сервису известно, для каждого набора входных данных можно определить, находился ли сервер в состоянии перегрузки или нет.

Набор данных для обучения представляет собой набор числовых значений, характеризующих выполнение одного запроса, и включает следующие сведения: время обработки запроса к заданному сервису (наличие перегрузки), количество запросов, находящихся в обработке в момент поступления данного запроса на сервер, значения индикативных величин, характеризующих состояние сервера в момент поступления данного запроса на сервер.

В дальнейшем для поиска первоначальных значений весов синаптических связей используется метод обучения с учителем на основе полученных обучающих примеров.

В целом применение предлагаемой методики идентификации перегрузки состоит в последовательном выполнении следующих действий:

Этап 1. Производится нагрузочное тестирование сервера, в ходе которого формируется выборка значений индикативных величин k_j ($j \in \bar{M}$), характеризующих состояние сервера, производится оценка наличия перегрузки для каждого случая обработки запроса в зависимости от заданного предельного времени его обработки.

Этап 2. Определяется множество индикативных величин k_j ($j \in M \subset \bar{M}$), в совокупности отражающих загруженность сервера, значения которых будут использоваться как входные данные для нейронной сети.

Этап 3. С использованием полученных обучающих примеров проводится первоначальный расчет синаптических весов нейронной сети w_{1j}, w_{2i}, например, при помощи метода обратного распространения ошибок.

Этап 4. При поступлении очередного запроса, если время обработки запроса t_i, рассчитанное на основе актуальных на момент поступления запроса входных данных нейронной сети, превышает значение L, запрос отклоняется, так как может вызвать перегрузку, иначе – принимается к обслуживанию.

Этап 5. Периодически с целью повышения точности идентификации перегрузки с учетом практических результатов управления запросами синаптические веса нейронной сети актуализируются.

Следует отметить, что этапы 1–3 являются подготовительными.

Для практической проверки предлагаемого подхода использовались данные, полученные по результатам нагрузочного тестирования web-сервера (конфигурация сервера: одноядерный процессор AMD Athlon64 3500+, 2 ГБ ОЗУ, 1 ТБ НЖМД, операционная система Windows Server 2008 R2, web-служба IIS), запросы к которому направлялись с одной рабочей станции (конфигурация рабочей станции: одноядерный процессор Intel Celeron 440M, 2 ГБ ОЗУ, 120 ГБ НЖМД, операционная система Linux Ubuntu 13.04) с использованием специализированного программного средства генерации потока запросов Tsung. Сервер и рабочая станция были соединены друг с другом по технологии Ethernet максимальной пропускной способностью 100 Мбит/сек. В ходе тестирования загруженность канала связи в среднем составляла менее 1 %, что позволяет не учитывать влияние фактора задержек по сети при оценке времени выполнения запросов.

В рамках web-сервера функционировали два различных сетевых сервиса, отличающиеся друг от друга по структуре ресурсоемкости.

Сетевой сервис № 1 выполнял операции по генерации псевдослучайных чисел программным способом. Структура ресурсоемкости была ориентирована на большое число вычислительных операций. В качестве ответа сетевой сервис возвращал уведомление об окончании выполнения операций.

Сетевой сервис № 2 выполнял операции по генерации псевдослучайных чисел программным способом и их последовательную запись в оперативную память. Структура ресурсоемкости была ориентирована на большое число вычислительных операций и активное использование оперативной памяти. В качестве ответа сетевой сервис возвращал уведомление об окончании выполнения операций.

Во время нагрузочного тестирования со стороны клиентского узла на web-сервер направлялись запросы к обоим сетевым сервисам с изменяемой интенсивностью таким образом, чтобы в различные моменты времени одновременно на сервере выполнялось различное число запросов к одному и другому сетевым сервисам.

Хотя в процессе тестирования были собраны сведения о запросах к обоим сетевым сервисам, при анализе данных и применении методики использовались данные о поступлении запросов к сетевому сервису № 2.

Для проверки предлагаемой методики идентификации собранные данные были использованы в качестве обучающих примеров нейронной сети. В качестве входных данных использовались показатели счетчиков производительности и количество запросов в обработке на сервере, в качестве выходных – фактическое время обработки запроса на сервере. Моделирование нейронной сети (схема сети 15–12–1) производилось для одного из двух сетевых сервисов с использованием ПО Statistica.

В качестве индикатора перегрузки принималось время обработки запроса, превышающее 40 секунд. В результате были получены сведения о выполнении 1783 запросов, из которых 769 были обработаны в режиме перегрузки. В результате нейронная сеть определила отсутствие перегрузки в 805 случаях, из них 45 ошибочно (5,59 %), наличие перегрузки – в 978 случаях, из них 254 ошибочно (25,97 %). В целом доля ошибок составила 16,77 %.

Для сравнения был построен прогноз возникновения перегрузки традиционным методом, идентифицирующим ее по факту возникновения. Суть метода заключается в том, что при поступлении очередного запроса прогноз относительно возникновения состояния перегрузки строится по результатам последнего обработанного запроса: если последний обработанный запрос не вызвал перегрузку, то и вновь поступивший также ее не вызовет, и наоборот. В результате отсутствие перегрузки было определено в 1066 случаях, из них 254 ошибочно (23,83 %), наличие перегрузки в 717 случаях, из них 202 ошибочно (28,17 %). В целом доля ошибок составила 25,57 %.

Полученные результаты показывают, что предложенный подход позволяет более точно предсказать состояние перегрузки, что дает возможность принимать обоснованное решение об обслуживании или отказе запросов и тем самым организовывать качественное управление запросами с учетом состояния сервера и текущей загруженности, учитывая индивидуальные особенности аппаратно-программной платформы web-сервера с целью предотвращения его перегрузки и обеспечения стабильности работы.

* Работа выполнена при поддержке Программы стратегического развития ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности на 2012–2016 гг.

СПИСОК ЛИТЕРАТУРЫ

1. Головко В. А., Безобразов С. В. Проектирование интеллектуальных систем обнаружения аномалий // Материалы международной научно-технической конференции OSTIS-2011. Минск: БГУИР, 2011.
2. Жуков А. В. Некоторые модели оптимального управления потоком входных заявок в интранет-системах // Новые информационные технологии в ЦБП и энергетике: Материалы 6-й научно-технической конференции. Петрозаводск, 2004. С. 87–90.
3. Кучерявый Е. А. Управление трафиком и качество обслуживания в сети Интернет. М.: Наука и техника, 2004.
4. Каллан Р. Основные концепции нейронных сетей: Пер. с англ. М.: Издательский дом «Вильямс», 2001. 290 с.
5. Менаске Д. А., Алмейда В. А. Ф. Производительность Web-служб. Анализ, оценка и планирование: Пер. с англ. СПб.: ООО «ДиаСофтЮП», 2003. 480 с.

Gusev O. V., Petrozavodsk State University (Petrozavodsk, Russian Federation)
Zhukov A. V., Petrozavodsk State University (Petrozavodsk, Russian Federation)
Peshkova I. V., Petrozavodsk State University (Petrozavodsk, Russian Federation)

WEB SERVER OVERLOAD IDENTIFICATION BY USE OF NEURAL NETWORK

Problems of diagnosing web servers' overloading are considered. It was found out that the use of simple models doesn't allow achievement of exact overload identification, but the use of complex models helps to assess the acceptable performance level of identification algorithm. When using a real Web server a set of features inherent to specific hardware-software platform can have a direct impact on its ability to resist overloads. The approach based on the server overload identification with the use of a neural network with the architecture of the three-layer direct distribution preceptor is offered. The range of the data source, necessary for construction of the set of input data of a neural network, is designated. The algorithm of actions for preparation of basic data, their analysis, and application of the received results for prediction of the origin of overloads on a Web server is offered. The offered approach is tested by data received during load testing of the Web server. The obtained results demonstrated that the offered approach is facilitative in a more precise overload assessment, which helps to organize effective requests' management with the purpose of congestion avoidance.

Key words: overload, neural networks, requests' managing, Web server, load testing, request

REFERENCES

1. Golovko V. A., Bezobrazov S. V. Design of intellectual systems of detection of anomalies [Proektirovanie intellektual'nykh sistem obnaruzheniya anomaliy]. *Materialy mezhdunarodnoy nauchno-tekhnicheskoy konferentsii OSTIS-2011* [Materials of the international scientific and technical OSTIS-2011 conference]. Minsk, BGUIR Publ., 2011.
2. Zhukov A. V. Some models of optimum control of a flow of input requests in the intranet systems [Nekotorye modeli optimal'nogo upravleniya potokom vkhodnykh zayavok v intranet-sistemakh]. *Novye informatsionnye tekhnologii v TsBP i energetike: Materialy 6-y nauchno-tekhnicheskoy konferentsii* [New Information Technologies in TsBP and Power Engineering: Materials of the 6th scientific and technical conference]. Petrozavodsk, 2004. P. 87–90.
3. Kucheryavyy E. A. *Upravlenie trafikom i kachestvo obsluzhivaniya v seti Internet* [Traffic management and quality of service on the Internet]. Moscow, Nauka i tekhnika Publ., 2004.
4. Kallan K. R. *Osnovnye kontseptsii neyronnykh setey* [The Essence of Neural Networks]. Moscow, Wil'yams Publ., 2001. 290 p.
5. Menasce D. A., Almeyda V. A. F. *Proizvoditel'nost' Web-sluzhb. Analiz, otsenka i planirovanie* [Capacity planning for web-services: methrics, models and methods]. St. Petersburg, OOO "DiaSoftYuP" Publ., 2003. 480 p.

Поступила в редакцию 06.11.2013