О ЗАДАЧЕ ВЫБОРА ОПТИМАЛЬНОЙ КОНФИГУРАЦИИ СИСТЕМЫ С СЕРВИСНО-ОРИЕНТИРОВАННОЙ АРХИТЕКТУРОЙ

Д. Душкин

Федеральное государственное бюджетное учреждение науки
Институт проблем управления
им. В. А. Трапезникова Российской академии наук
Москва, Россия
ddushkin@asmon.ru

В работе представлен краткий обзор современного состояния области моделирования сервисно-ориентированной архитектуры и веб-сервисов. Рассмотрен вопрос оптимального выбора веб-сервисов по предпочтению, используемых в системах с сервисно-ориентированной архитектурой. Описан новый критерий сравнения веб-сервисов, позволяющий оценить возможность обеспечения определенного уровня производительности при возрастающей нагрузке, – чувствительность.

Ключевые слова: веб-сервисы, сервисно-ориентированная архитектура, машинное обучение, анализ чувствительности.

1. ВВЕДЕНИЕ

Предоставление сервиса вместо аппаратного и программного обеспечения стало одним из ведущих направлений современной индустрии информационных технологий. Данная тенденция во многом меняет экономику сферы информационных технологий и оказывает влияние на формирование информационного общества в целом [1].

В основе сервис-ориентированных систем лежит *веб-сервис* – автономный вычислительный ресурс, предоставляющий свои функции через сеть Интернет (или Интранет) посредством открытых протоколов обмена данными, не зависящих от платформ как самих ресурсов, так и связываемых с ними программных систем. Системы, использующие в своей работе веб-сервисы, называют системами с сервисно-ориентированной архитектурой (СОА).

Часто одну и ту же функцию предоставляют различные сервисы. Например, функцию катографии предоставляют сервисы Google Maps, Яндекс Карты, Bing Maps (сервис Microsoft), Nokia Maps и другие. В таком случае перед инженерами, проектирующими систему с СОА, ставится многокритериальная задача выбора оптимального по предпочтениям веб-сервиса, предоставляющего функцию картографии. В настоящей работе представлен краткий обзор современного состояния области моделирования сервисно-ориентированной архитектуры и веб-сервисов. Описан новый критерий сравнения веб-сервисов, характеризующий производительность, – чувствительность.

В заключении сделаны выводы по проведенной работе и обозначены предстоящие направления исследований.

2. ОБЗОР

В российской научной литературе можно выделить всего несколько работ, касающихся исследования веб-сервисов и разработки алгоритмов их оптимального выбора. В работе [2] веб-сервисы используются как ресурсы, содержащие различные образовательные материалы. В качестве критериев сравнения выбраны класс онтологии и степень близости ресурсов веб-сервиса к заданному классу онтологии.

В работе [3] предлагается алгоритм выбора резервных веб-сервисов на основе вычисления интегральной оценки веб-сервиса – весовой функции, вычисляемой по следующим критериям: количество отказов, деленное на количество вызовов (другими словами доступность), среднее время обслуживания запроса и стоимость одного запроса, заданная экспертом.

В работе [4] рассматривается применение теории нечетких чисел в задаче выбора сервисов для реализации определенных бизнес-процессов в рамках корпоративной информационной системы. Используя терминологию настоящей работы, можно сказать, что в [4] выделяют ряд функциональных критериев (критерии реализации бизнес-процессов) и нефункциональных. К нефункциональным относятся оценки экономических затрат: единовременных, периодических и косвенных. Другие нефункциональные критерии не рассматриваются.

Большое количество исследований по сервисно-ориентированным вычислениям проведено в рамках европейского научно-исследовательского проекта SENSORIA. Часть результатов этих исследований посвящена методам оценки качества веб-сервисов (см., например, [5–11]). Необходимым условием применимости данных методов является наличие модели оцениваемого веб-сервиса в формализме одной из алгебр случайных процессов. Разработка такой модели требует определенных затрат времени и ресурсов, которые имеет смысл затрачивать разработчику веб-сервиса, но не внешнему пользователю, периодически сталкивающемуся с необходимостью выбора одного из существующих на рынке веб-сервисов для реализации одной из многих функций разрабатываемой им системы массового обслуживания.

Из обзора можно сделать вывод о малом внимании к критериям, описывающим производительность и надежность веб-сервисов при растущей нагрузке. Такие сведения важны при проектированнии системы с СОА на этапе определения максимальной загруженности, при которой система сможет обеспечить приемлемый уровень предоставляемых услуг при данном наборе используемых веб-сервисов. В настоящей работе исследуется критерий чувствительности, позволяющий решить описываемую проблему.

3. ОПРЕДЕЛЕНИЕ ЧУВСТВИТЕЛЬНОСТИ ВЕБ-СЕРВИСА

Чувствительность веб-сервиса – критерий, на основе которого может быть оценена возможность обеспечения определенного уровня производительности веб-сервиса при возрастающей нагрузке. Выделения классов чувствительности веб-сервисов реализуется следующим образом:

1) Формируется случайная выборка веб-сервисов (см. подраздел 3.1).
2) Проводится тестирование выборки по определенному плану (см. подраздел 3.2).
3) Составляется матрица «объект-признак». С этой целью полученные данные преобразовываются для выделения ряда признаков (факторов), отражающих характер изменения производительности при возрастающей нагрузке (см. подраздел 3.3).
4) Проводится эвристическое разделение множества веб-сервисов на классы по уровню чувствительности, экспертно классы упорядочиваются по предпочтению (см. подраздел 3.4).
5) Для проверки корректности эвристического разделения веб-сервисов на классы проводится кластеризация данных с помощью алгоритма машинного обучения «без учителя» – методом k-средних (см. подраздел 3.5).

3.1. Формирование выборки. Выборка веб-сервисов формируется на основе данных из каталога API Directory, содержащего информацию о более чем 5000 различных веб-сервисах. Выбираются разнородные сервисы, реализующие функции картографии и геокодинга, предоставляющие информацию о различных показателях торговых бирж, о погоде, новостях и т.д. Все сервисы предоставляют свои функции по протоколу REST.

3.2. План теста. Задача теста – определить величину среднего времени обработки запросов, стандартное отклонение времени обработки запросов и количество необработанных запросов при заданной нагрузке. В процессе тестирования осуществляется последовательное выполнение итераций, отличающихся числом запросов в секунду. В рамках каждой итерации в течение секунды отправляется определенное число запросов к веб-сервису по протоколу HTTP. Запросы равномерно распределены в рамках секунды.

Пусть λ_{max} – максимальное число запросов в секунду, S – шаг теста (число, на которое увеличивается количество запросов в секунду в последующей итерации). Тогда можно посчитать общее число итераций в тесте N_{iter}:

$$N_{iter} = \left\lceil \frac{\lambda_{max}}{S} \right\rceil$$

Пусть $\mathbf{r} \in \mathbb{R}^k$ – вектор, содержащий время обработки запросов всего теста, где k – общее число отосланных запросов за все итерации. Для простоты дальнейших вычислений, если i-ый запрос не был обработан или был обработан с ошибкой, то $r_i = 0$[1].

[1]При вычислении среднего арифметического и стандартного отклонения исключаются те запросы, время обработки которых равно нулю ($r_i = 0$).

Пусть $Iter \in \mathbb{R}^{N_{iter} \times k}$ – матрица, содержащая информацию о принадлежности запросов к определенным итерациям. $Iter_{i,j} = 1$, если запрос j выполнялся в рамках итерации i, иначе $Iter_{i,j} = 0$.

3.3. Получение матрицы «объект-признак». Матрица «объект-признак» $X \in \mathbb{R}^{m \times n}$, где m – число объектов, n – количество признаков, формируется на основе результатов тестов веб-сервисов. Результат теста состоит из вектора времен обработки запросов $\mathbf{r}$ и матрицы распределения запросов по итерациям $Iter$. Ряды матрицы соответствуют объектам (веб-сервисам), столбцы – признакам. Обозначим $\mathbf{x}^{(i)}$ i-ый ряд матрицы X. Вектор $\mathbf{x}^{(i)}$ является объединением трех векторов: $\mathbf{x}^{(i)} = \mathbf{t}^{(\mathbf{i})} \cup \mathbf{d}^{(\mathbf{i})} \cup \mathbf{e}^{(\mathbf{i})}$, где вектор $\mathbf{t}^{(i)} \in \mathbb{R}^{N_{iter}}$ содержит среднее время обработки запросов по итерациям i-го веб-сервиса, $\mathbf{d}^{(i)} \in \mathbb{R}^{N_{iter}}$ – стандартное отклонение времен обработки запросов по итерациям, $\mathbf{e}^{(i)} \in \mathbb{R}^{N_{iter}}$ – количество ошибочных или необработанных запросов по итерациям (далее для краткости такие запросы будем называть просто «необработанные запросы»). Таким образом общее количество признаков $n = 3 * N_{iter}$.

Вектор $\mathbf{t}^{(i)}$ вычисляется следующим образом:

$$t_j^{(i)} = \frac{1}{j \cdot S} \sum_{l:Iter_{j,l}^{(i)}=1} r_l^{(i)}$$

где $i = \overline{1, ..., n}$ – номер веб-сервиса, $j = \overline{1, ..., N_{iter}}$ – номер итерации, $l = \overline{1, ..., k}$ – номер запроса, $\mathbf{r}^{(i)}$ – вектор, содержащий время обработки запросов i-го веб-сервиса, $Iter^{(i)}$ – матрица, содержащая информацию о принадлежности запросов к определенным итерациям. Запись $\sum_{l:Iter_{j,l}^{(i)}=1} r_l^{(i)}$ следует понимать как «сумма по элементам $r_l^{(i)}$, где l такое, что $Iter_{j,l}^{(i)} = 1$», или «сумма по элементам, принадлежащим i-ой итерации».

Вектор $\mathbf{d}^{(i)}$:

$$d_j^{(i)} = \sqrt{\frac{1}{j \cdot S - 1} \sum_{l:Iter_{j,l}^{(i)}=1} (r_l^{(i)} - t_j^{(i)})^2}$$

Вектор $\mathbf{e}^{(i)}$:

$$e_j^{(i)} = \sum_{l:Iter_{j,l}^{(i)}=1, r_l=0} 1$$

Нормализуем полученные данные. Для этого зададим функцию нормализации g_j:

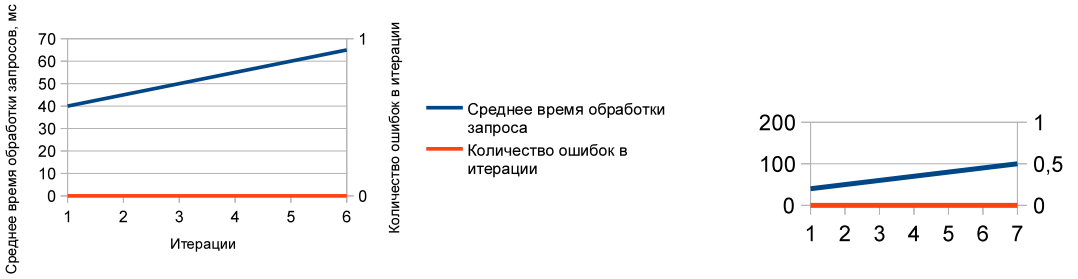
$$g_j : x_{i,j} \mapsto \frac{x_{i,j} - \min_j x_{i,j}}{\min_j x_{i,j} - \min_j x_{i,j}}$$

Примем за $X' = g(X)$ нормализованную по признакам матрицу «объект-значение».

3.4. Эвристическое выделение классов чувствительности веб-сервиса. По результатам проведенных тестов было выявлено, что реальные веб-сервисы имеют большой разброс значений критериев. Анализируя полученные массивы данных, нетрудно определить случаи низкой чувствительности, когда повышение нагрузки практически

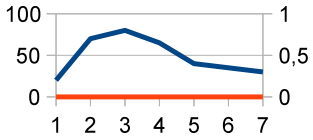
не влияет на значения показателей, и случаи высокой чувствительности, когда небольшое повышение нагрузки значительно увеличивает среднее время обработки запросов и часто ведет к отказу в обслуживании большей части запросов. Однако большинство веб-сервисов демонстрируют промежуточное поведение и возникают сложности с определением их класса чувствительности.

Ниже представлено эвристическое разделение веб-сервисов на классы чувствительности: от предпочтительной низкой чувствительности к высокой.

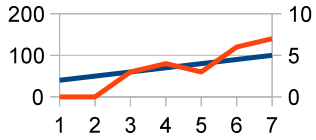


(a) 1 класс чувствительности (низкая чувствительность)

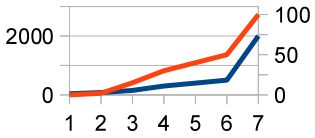
(b) 2 класс чувствительности



(c) 3 класс чувствительности



(d) 4 класс чувствительности



(e) 5 класс чувствительности (высокая чувствительность)

Рис. 1. Концептуальное определение классов чувствительности

Первый класс (рис. 1(a)) – низкая чувствительность, характеризуется медленным повышением (иногда отсутствием повышения) среднего времени обслуживания запросов, отсутствием необработанных запросов.

Второй класс (рис. 1(b)) – средняя чувствительность, характеризуется более быстрым повышением среднего времени обслуживания запросов по сравнению с первым классом, отсутствием необработанных запросов.

Третий класс (рис. 1(c)) – повышение с последующим понижением среднего времени обслуживания запросов, отсутствие необработанных запросов. Такое поведение характерно для ряда «облачных» веб-сервисов (при увеличении нагрузки динамически увеличивается мощность обслуживающего узла) или веб-сервисов с адаптивным распределителем нагрузки (англ. *Load balancer*) (при высокой утилизации ресурсов одного обслуживающего узла часть запросов передается на обслуживание узлам с меньшей утилизацией).

Четвертый класс (рис. 1(d)) по характеру повышения среднего времени обслуживания похож на второй, но с повышением нагрузки появляется небольшое число необработанных запросов.

Пятый класс (рис. 1(e)) – высокая чувствительность, характеризуется резким повышением среднего времени обслуживания, большим числом необработанных запросов.

3.5. Автоматизированное выделение классов чувствительности. С целью проверки обоснованности эвристического разделения веб-сервисов, а также с учетом последующей автоматизации процесса определения класса чувствительности используется кластеризация данных с помощью алгоритма k-средних. В качестве входных данных используются нормализованная матрица «объект-признак», 5 центройдов (по количеству эвристически определенных классов), каждую итерацию центройды выбираются случайно, всего проводится 50 итераций, в качестве меры расстояния используется расстояния Евклида.

4. ПРОВЕДЕНИЕ ТЕСТОВ И АНАЛИЗ ПОЛУЧЕННЫХ ДАННЫХ

Зададим план теста: максимальное число запросов в секунду $\lambda_{max} = 300$, шаг теста $S = 10$. Такие значения параметров выбраны экспериментально, т.к. было установлено, что начиная с 280-290 запросов в секунду большинство веб-сервисов демонстрируют устойчивое поведение. Выборка состоит из 50 веб-сервисов: Google Maps, Яндекс Карты, Bing Maps, Nokia Maps, Twitter, Factolex, Quora и др. Исходя из плана теста несложно определить размер матрицы $X \in \mathbb{R}^{50\times 90}$, т.е. матрица представляет 50 объектов, каждый из которых характеризуется 90 показателями.

На рис. 2 показана первая группа признаков «среднее время отклика по итерациям» 5-ти веб-сервисов (если отобразить на графике все 50 веб-сервисов, то график станет слишком зашумленным). Данные выбраны таким образом, чтобы веб-сервисы наглядно можно было бы отнести к различным классам. Каждая линия соответствует одному веб-сервису.

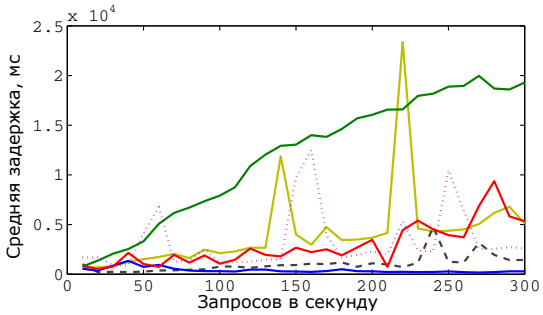
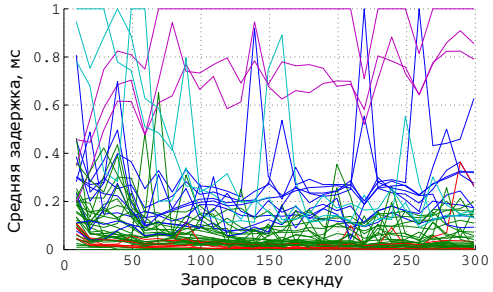


Рис. 2. Ненормализованные группы признаков

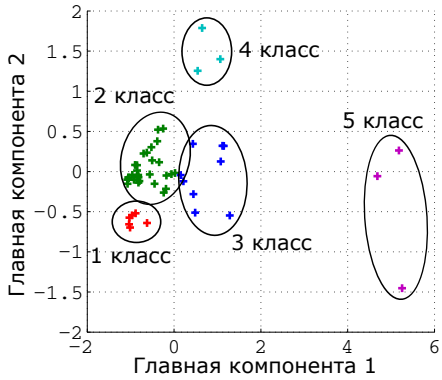
Результаты кластеризации можно визуализировать с помощью метода главных компонент, уменьшив размерность с 90 до 2 признаков, наиболее полно объясняющих изменчивость и взаимосвязи исходных данных. На графике каждый плюс обозначает один веб-сервис; плюсы, закрашенные одним цветом, принадлежат одному кластеру.

На рис. 3(b) хорошо видна разделяемость множества веб-сервисов на различные кластеры. Сопоставляя рис. 3(b) и рис. 3(a), а также руководствуясь эвристической

классификацией, описанной в подразделе 3.4, можно соотнести кластеры с классами чувствительности.



(a) Нормализованная группа признаков «среднее время отклика от числа запросов в секунду» с распределенными по кластерам веб-сервисами



(b) Применение метода главных компонент к результатам кластерного анализа

Рис. 3. Результаты кластеризации

5. ЗАКЛЮЧЕНИЕ

На основе проведенных исследований можно сделать вывод – реальные веб-сервисы демонстрируют различную производительность и надежность при повышении нагрузки. Выявлению такого показателя, который смог бы охарактеризовать различное поведение веб-сервисов при повышении нагрузки, посвящена настоящая работа.

Используя корректное программное обеспечение для проведения тестов и сбора данных, методы анализа данных и машинного обучения возможно автоматизировать процесс определения класса чувствительности веб-сервиса. Автоматизация такого процесса позволит учитывать перспективную производительность и надежность веб-сервисов на этапе проектирования систем с сервисно-ориентированной архитектурой, что особенно полезно при создании систем с высокой нагрузкой.

Видится перспективным продолжение исследований по данной тематике. Возможно выделение новых признаков (и отбор наиболее значимых из имеющихся), характеризующих чувствительность. Полезна формализация плана проводимого теста с учетом закона распределения среднего времени обработки запросов и погрешностей, возникающих вследствие неизвестного внешнего трафика веб-сервиса. Используя аппарат теории принятия решений, возможно разработать алгоритмы выбора оптимальной конфигурации систем с сервисно-ориентированной архитектурой.

ЛИТЕРАТУРА

1. *M. Wirsing, D. N. Rocco, and S. Gilmore.* SENSORIA process calculi for service-oriented computing // Lecture Notes in Computer Science. Springer. 2007. V. 4661.

2. *Смирнов А. В., Левашова Т. В., Шилов Н. Г.* Конфигурирование сервис-ориентированных сетей ресурсов для интеллектуальной поддержки дистанционного образования // Открытое образование. 2010. V. 2. P. 111–117.
3. *Бабошин А. А., Кашевник А. М.* Подход к организации взаимодействия веб-сервисов на основе модели потока работ // Труды СПИИРАН. 2007. V. 5. P. 247–254.
4. *Затеса А. В.* Подход к организации взаимодействия веб-сервисов на основе модели потока работ // Труды СПИИРАН. 2007. V. 5. P. 247–254.
5. *Brewer, Eric A.* Towards robust distributed systems // Proceedings of the nineteenth annual ACM symposium on Principles of distributed computing. 2000. V. 19
6. *I. Cappello, A. Clark, S. Gilmore, D. Latella, M. Loreti, P. Quaglia, and S. Schivo* Quantitative analysis of services // Rigorous software engineering for service-oriented systems. Springer-Verlag. 2011. P. 522–540.
7. *H. Foster, L. Gönczy, N. Koch, P. Mayer, C. Montangero, and D. Varró* Uml extensions for service-oriented systems // Rigorous software engineering for service-oriented systems. Springer-Verlag. 2011. P. 35–60.
8. *S. Gilmore, L. Gönczy, N. Koch, P. Mayer, M. Tribastone, and D. Varró* Non-functional properties in the model-driven development of service-oriented systems // Software Systems Modelling 2011. V. 10. P. 287–311.
9. *E. M. Maximilien and M. P. Singh* A framework and ontology for dynamic web services selection // IEEE Internet Computing. 2005. V. 8. P. 84–93.
10. Tribastone M., Gilmore S. Scaling performance analysis using fluid-flow approximation // Rigorous software engineering for service-oriented systems. Springer-Verlag. 2011. P. 486–505.
11. Wu Q., Iyengar A., Subramanian R., Rouvellou I., Silva-Lepe I., Mikalsen T. Combining quality of service and social information for ranking services // Service-Oriented Computing. Springer Berlin Heidelberg. 2009. V. 5900. P. 561–575.