Модели оптимизации функционирования Информационного веб-портала

А. В. Босов, А. В. Борисов

Институт проблем информатики РАН,
Россия, 119333 Москва, ул. Вавилова, 44, корп. 2

Рассматриваются задачи оценивания двух показателей функционирования веб-портала — показателя эффективности информационных источников и показателя пользовательской активности. Для анализа показателей сформулированы и решены задачи оценивания состояний динамических систем наблюдения. Предложенные математические модели применяются для оптимизации работы основного интеграционного компонента веб-портала. Обсуждается методика определения параметров моделей.

Введение

Одной из задач, возникающих при разработке современных информационных систем, является задача интеграции разнородных источников информации. Для ее решения в распределенной среде, формируемой множеством взаимодействующих систем, применяются разные инструменты, к которым, в том числе, относятся и веб-порталы [1, 2]. Множество деталей, свойственных портальной тематике, для целей данной работы не важны, поскольку предметом исследования является конкретный программный продукт — Информационный веб-портал [3]. Его базовое функциональное назначение состоит в создании среды для интеграции разнородных информационных систем («внешних» информационных источников), объединяемых на федеративной основе. Веб-портал, таким образом, рассматривается в качестве средства для создания центрального узла распределенной федеративной среды, построенной на стандартах Интернета.

В архитектуре Информационного веб-портала [4] присутствуют традиционные для веб-систем элементы, такие как подсистема управления содержанием [5], подсистема безопасности [6], подсистема доступа к данным [7]. Ключевым же элементом, обеспечивающим базовую функциональ-

ность в задаче интеграции, является подсистема интеграции и поиска [8]. Данная подсистема не только отвечает за поддержку ключевой портальной функциональности, но и является тем местом, от функционирования которого зависит в наибольшей степени общая оценка качества всего программного решения в целом. Более того, оказывается, что алгоритм работы именно этой подсистемы является наиболее вариативным — допускает различные формы настройки, изменения параметров и проч., поэтому оптимизации его работы и уделяется наибольшее внимание.

Аппарат, который может быть применен с указанной целью, может основываться на использовании самых разных математических методов. Для выбора подходящего математического аппарата следует учесть, что рассматриваемая веб-система, хотя и определена довольно формально, и работает по фиксированным (детерминированным) алгоритмам, однако пребывает в окружении не просто неизвестных или неконтролируемых факторов, а входит в состав довольно сложной конструкции — распределенной аппаратно-программной среды, обслуживающей множество функциональных запросов множества пользователей. Для портала эта среда выглядит, по меньшей мере, не вполне определенной. Так, например, средствами портала не могут контролироваться состояния информационных источников, с которыми он взаимодействует, текущая нагрузка на телекоммуникационную составляющую, которая используется не только портальными приложениями, но и массой других. Наибольший вклад в упомянутую неопределенность вносит поведение пользователей, которые меняют интенсивность работы в зависимости от своих текущих задач, меняют интересы и степень востребованности элементов доступного через портал контента и т. п. Таким образом, предложить для описания функционирования портала адекватную математическую модель без учета свойственных системе неопределенностей вряд ли возможно. Описывать же потребности пользователей, как и прочие неконтролируемые порталом, но влияющие на его работу факторы, следует с учетом их случайного характера. По этой причине представляется вполне обоснованным применять в этом случае математический аппарат теории стохастических систем [9]. Возможности моделирования, доступные в рамках этой теории, достаточно богаты, а развитость аппарата позволяет получать окончательные решения для широкого класса задач оптимизации.

Постановки задач оптимизации процессов функционирования Информационного веб-портала, в частности рассматриваемых далее задач оптимального оценивания показателей эффективности взаимодействующих с порталом информационных источников и показателей пользовательской активности, опираются на некоторый общий подход к моделированию. В разных случаях для решения конкретной задачи оптимизации используется некоторый показатель функционирования портала. Это могут быть

временные и объемные характеристики взаимодействия портала с «внешними» информационными источниками или собственным хранилищем, показатели пользовательской активности, характеристики выделяемых и/или требующихся вычислительных ресурсов и т. п. Предложить модель непосредственно из физического смысла показателя, как правило, оказывается затруднительным, поэтому модель следует идентифицировать. Использование традиционных линейных моделей идентификации [10] представляется малоперспективным: как правило, в динамике рассматриваемых показателей легко просматриваются черты нелинейных систем, такие как зависимость возмущений от текущих значений показателя, цикличность процессов, скачкообразные изменения характеристик и проч. Для учета таких явлений, с одной стороны, и для возможности использования в решениях полезных свойств линейных систем — с другой, предлагается подход на основе классификации возможных состояний изучаемого показателя. Пространство значений показателя разбивается на области, и предполагается, что для значений показателя внутри области его динамика описывается простейшими линейными уравнениями, а при выходе показателя за границы области модель изменяется. Как правило, значения рассматриваемого показателя прямым измерениям недоступны, поэтому формулируется задача оценивания показателя по результатам косвенных наблюдений. Решение задачи оценивания используется далее для выработки обоснованного управляющего воздействия, зависящего от значений моделируемого показателя.

Алгоритм работы подсистемы интеграции и поиска Информационного веб-портала, как уже упоминалось, зависит от нескольких таких показателей. В данной работе рассмотрены два ключевых — показатель эффективности взаимодействия информационных источников портала и показатель пользовательской активности.

1. Описание задачи управления пулом запросов

Возможность оптимизации работы подсистемы интеграции и поиска изыскивается в самом алгоритме ее работы при обработке пользовательского запроса, направляемого для выполнения несколькими информационными источниками, подключенными к порталу. Данный алгоритм состоит в выполнении следующих действий:

1) пользователь выбирает тип ресурса для поиска, заполняет поля поисковой формы, устанавливает иные атрибуты поиска;
2) пользовательский запрос, ассоциированный с поисковой командой (под командой понимается формальное задание критериев отбора данных и перечень источников, которые должны предоставлять информацию),

описание которой имеется в схеме портала, принимает подсистема управления содержанием, опознает команду и передает ее на выполнение подсистеме интеграции и поиска;

3) выполняется анализ метаданных подключенных источников и выявляются источники, поддерживающие данный тип ресурса; вместе с этим извлекаются описание команды и иная служебная информация, необходимая для формирования запросов в терминах схем источников и активизации их адаптеров;
4) полученное множество запросов (число запросов равно числу источников, поддерживающих выбранный пользователем тип с учетом возможных указаний пользователя опрашивать не все источники) направляется компоненту выполнения запросов; *данный компонент распределяет запросы по нитям из поддерживаемого им пула запросов*;
5) информация, полученная из источников, консолидируется по мере выполнения запросов; после получения результата выполнения последнего запроса окончательный результат направляется подсистеме представления, обеспечивающей доведение результатов пользователю (имеется также принципиальная возможность направлять пользователю результаты по мере их поступления, но окончательный результат все равно оформится только по получении последнего ответа).

Для наглядности описанный процесс проиллюстрирован на рис. 1.

Местом для оптимизации является шаг 4, на котором выполняется распределение подготовленных L_{II} запросов (в общем случае максимальное число запросов совпадает с числом источников; так оказывается, например, если пользователем задана команда на полнотекстовой поиск без ограничения участвующих источников, а каждый из источников поддерживает такую функцию) по имеющемуся у компонента выполнения пулу. Пул представляет собой набор заранее инициализированных нитей (здесь будем считать, что размер пула определяется при настройке портала и не меняется в процессе работы, а из общего числа L_{Π} нитей пула на выполнение текущей команды выделено $L_{H\Pi}$ нитей), каждой нитью поддерживается очередь, в которую направляются запросы и необходимая служебная информация для активизации адаптера соответствующего источника. В каждой нити выбирается следующий запрос из очереди, определяется нужный адаптер, выполняется его вызов и прием результата выполнения запроса. Распределение запросов к источникам по пулу определяет суммарное время, затрачиваемое на выполнение пользовательского запроса, т. е. формирует основной вклад в субъективную оценку эффективности портала. По этой причине распределять запросы по нитям надо некоторым практичным образом, с тем чтобы эффективность портала оценивалась как можно выше.

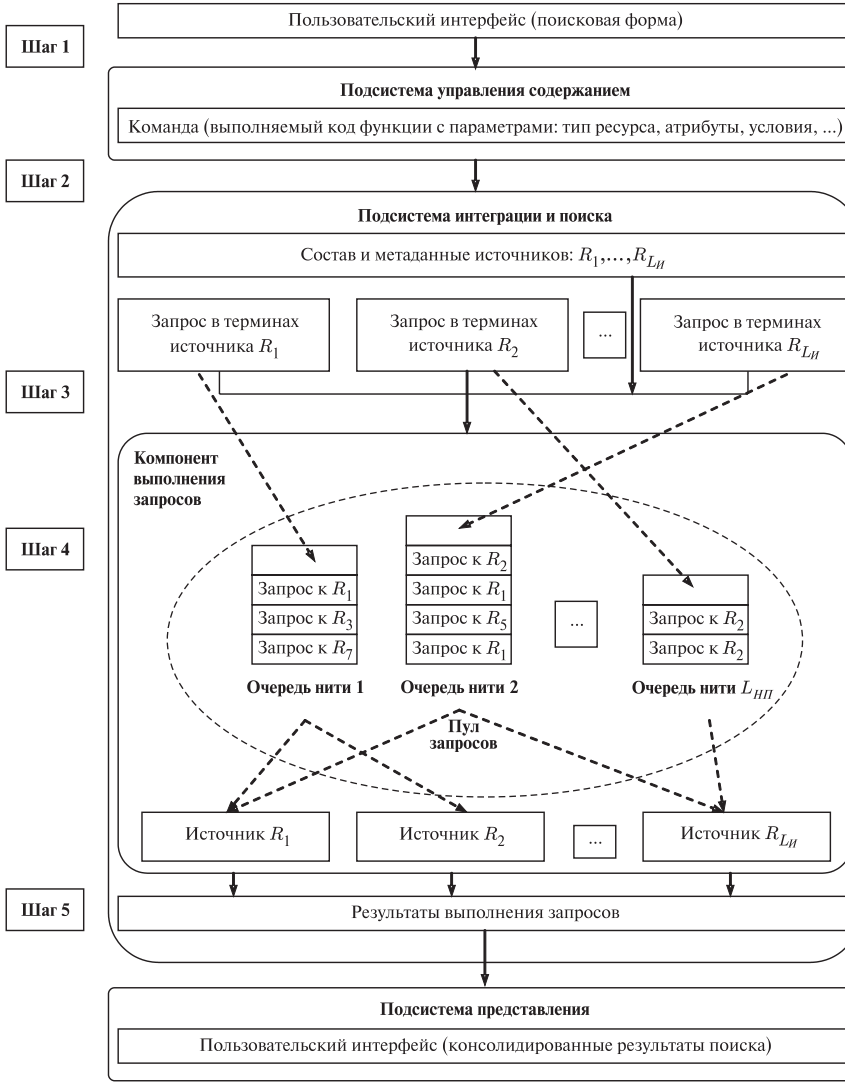

Рис. 1. Алгоритм работы подсистемы интеграции и поиска

Таким образом, формируется первая задача оптимизации функционирования портала, состоящая в оптимальном распределении множества запросов по пулу (*по множеству очередей*) *с целью минимизации затрат на суммарное время выполнения пользовательской команды* (*множества запросов*).

Может показаться, что распределять запросы возможно гораздо проще. Например, создать общую очередь, из которой последовательно извлекать запросы и направлять их очередной освободившейся нити. В этом случае вероятность неэффективного распределения была бы минимальной, и потребность в дополнительной оптимизации отсутствовала. Поступить так не удается, поскольку портал предполагает одновременную работу с множеством пользователей, а такая «последовательная» схема обработки означала бы и последовательное выполнение поступающих параллельно пользовательских команд. При этом пользователь, сформировавший «простой» запрос к одному источнику в то время, когда пул занят выполнением «сложных» запросов другого пользователя ко многим источникам, будет находиться в ожидании результата необоснованно долго. Избежать такой ситуации можно только ограничением числа нитей, отводимых под выполнение команды одного пользователя, т. е., как и обозначено выше, отводить под очередную команду $L_{НП}$ нитей — часть из общего размера пула $L_{П}$. Само по себе такое ограничение (определение числа $L_{НП}$) требует анализа и оптимизации, так как, очевидно, тоже является задачей управления пулом запросов (эта задача сводится к задаче анализа пользовательской активности).

Таким образом, формируется вторая задача оптимизации функционирования портала, состоящая в определении числа нитей (*части общего пула*)*, отводимых под выполнение одной пользовательской команды на основании анализа текущей пользовательской активности.*

В следующем пункте рассматривается первая из задач управления пулом в предположении, что число нитей $L_{НП}$, выделенных пользователю, сформировавшему очередную команду, фиксировано. Как определяется это число, т. е. как решается вторая задача управления пулом, обсуждается далее в раздее 2.

Поскольку кроме схемы источника и служебных метаданных никаких иных достоверных сведений об источнике нет, то оптимизировать распределение запросов по пулу возможно только на основании косвенной информации, т. е. наблюдений, формируемых в процессе взаимодействия портала с источником. Какие бы ни были наблюдения, в них обязательно будет присутствовать «вклад» от целого набора неконтролируемых факторов, среди которых, например, такие:

- различная производительность аппаратно-программного обеспечения, которое используют источники;
- разница в объемных характеристиках результатов выполнения запросов, зависящая не только от семантики конкретного запроса, но и от полноты данных источника;
- различная доступность источников, связанная с федеративным характером обслуживания и эксплуатации, и проч.

Перечисленные факторы с некоторой степенью идеализации могут быть охарактеризованы как статические, но есть и группа очевидно динамических факторов, влияющих на исследуемый процесс:

- изменения, внесенные в ресурсы источника (добавление, удаление, модификация экземпляров ресурсов);
- различный уровень нагрузки на аппаратное и системное программное обеспечение со стороны других приложений, существующих на той же платформе;
- различная пропускная способность сегментов сети, используемой при взаимодействии портала с источником;
- пользовательские предпочтения при выборе перечня источников, участвующих в выполнении запроса, и проч.

Собственно алгоритм распределения запросов по пулу максимально прост, поскольку ресурсоемкий алгоритм управления на этом шаге обработки команды приведет только к лишним затратам портальных ресурсов, которые нивелируют эффективность процедуры оптимизации. Именно, выполняется следующее. К служебным метаданным каждого информационного источника добавлена некоторая характеризующая его оценка (смысл этой оценки обсуждается далее). Периодически значение этой оценки изменяется на основании анализа текущего взаимодействия с источником. При обработке очередной пользовательской команды источники упорядочиваются так, что ключом является текущее значение этой оценки. Запросы распределяются по нитям пула запросов последовательно: первый запрос направляется в очередь первой нити, второй — в очередь второй нити и т. д., по исчерпании пула очередной запрос вновь направляется в очередь первой нити и т. д. Для того чтобы такой подход был эффективен, требуется, чтобы оценка, данная источнику, характеризовала текущие временные затраты на взаимодействие с ним портала, т. е. обеспечивала равномерное распределение «медленных» и «быстрых» запросов по нитям пула. Тогда «медленные» запросы будут распределены по разным нитям и, таким образом, окажут минимально возможное влияние на время ожидания выполнения пользовательской команды.

Требуемая оценка источника должна, с одной стороны, отражать его «потребительские» свойства (время, затрачиваемое на ожидание ответа), с другой — учитывать предыдущий опыт взаимодействия с ним, так как вполне может оказаться, что группа «медленных» источников в течение некоторого времени окажется просто невостребованной пользователями, что при возникновении интереса к ним приведет к существенной задержке, так как «потребительские» свойства при отсутствии взаимодействия будут нулевыми. По этой причине, в частности, не представляется целесообразным и привлечение эксперта, который мог бы назначить эти оценки волюнтаристски и не изменять их динамически.

Получить искомые оценки хотелось бы, основываясь на адекватной математической модели, описывающей показатель того, насколько каждый конкретный источник эффективен (требует меньше временных затрат в обслуживании запросов) в данный момент времени. Доступные для построения такой модели наблюдения должны формироваться за счет анализа результатов выполнения запросов данным источником. Естественно в качестве таких наблюдений выбрать время ожидания ответа на запрос (время выполнения запроса). Будем исходить из случайной природы этих наблюдений, учитывая тем самым множественный и неконтролируемый характер перечисленных выше факторов неопределенности.

Прежде чем перейти собственно к построению модели описанного явления, особо отметим, что при построении этой модели помимо случайного характера нужно учесть также, что:

- формируемая модель должна быть динамической, т. е. учитывать зависимость от времени «потребительских» свойств источников;
- для моделирования следует использовать дискретное время — задать интервалы наблюдения так, чтобы вычислять требуемую оценку только в отдельные моменты времени, расположенные друг от друга на достаточном расстоянии, с тем чтобы не тратить слишком много вычислительных ресурсов на процедуры оптимизации, нанося значительный ущерб собственной функциональности портала;
- число нитей в пуле запросов компонента выполнения фиксировано и не может быть динамически изменено, каждая нить инициализирована (под нее выделена память, создана очередь), и применение нити не требует дополнительных затрат.

2. Оценка эффективности источника

2.1. Модель формирования оценки эффективности

Предложить уравнения для показателя эффективности источника (например, для среднего времени выполнения запроса на текущем интервале наблюдения) непосредственно из физического смысла представляется затруднительным. По этой причине предлагается ввести некоторые условные эффективности путем задания классов, к которым относятся источники с точки зрения временных затрат на взаимодействие с порталом. Например, можно определить три класса:

- $\theta_Б$ — класс эффективных источников («быстрых»);
- θ_C — класс средне эффективных источников («средних»);
- θ_M — класс неэффективных источников («медленных»).

Тогда задача оптимизации распределения запросов по пулу сводится к задаче классификации на основании накапливаемых данных наблюдений (как предыстории взаимодействия, так и текущих наблюдений) информационных источников, т. е. к обоснованному отнесению конкретного источника $\Re$ к одному из классов $\theta_Б$, θ_C или θ_M.

Модель наблюдений должна описывать зависимость измеренных характеристик (время выполнения запроса) для источника $\Re$ от того, к какому классу он относится. В предположении, что эта зависимость линейна, уравнение для измеряемой в момент времени t характеристики примет вид

$$Y_t = \mathbf{C} \cdot \begin{pmatrix} I(\Re \in \theta_Б) \\ I(\Re \in \theta_C) \\ I(\Re \in \theta_M) \end{pmatrix} + \mathbf{\Sigma} \cdot \begin{pmatrix} I(\Re \in \theta_Б) \\ I(\Re \in \theta_C) \\ I(\Re \in \theta_M) \end{pmatrix} \cdot W_t, \qquad (1)$$

где Y_t – текущее наблюдение за источником $\Re$ (на практике измеряется среднее или суммарное время ожидания отклика источника за последний интервал наблюдения); $\mathbf{C}=(C_Б, C_C, C_M)$ — вектор-строка, числа $C_Б, C_C, C_M$ определяют среднюю (ожидаемую) величину наблюдения (среднее ожидаемое время отклика источника) в зависимости от класса, к которому отнесен источник $\Re$; W_t — возмущение (шум), моделирующее влияние на наблюдение перечисленных выше факторов неопределенности; $\mathbf{\Sigma}=(\Sigma_Б, \Sigma_C, \Sigma_M)$ — вектор-строка, числа $\Sigma_Б, \Sigma_C, \Sigma_M$ определяют отклонения наблюдаемых значений от средних величин в зависимости от класса, к которому отнесен источник $\Re$. Индикаторные функции $I(\Re \in \theta)$ в (1) принимают значение 1, если источник $\Re$ относится к классу θ, и 0 противном случае.

Практическое вычисление наблюдения, описываемого моделью (1), осуществляется следующим образом. Интервал наблюдения (работы портала) $[t_0; +\infty)$ разбивается на отрезки (можно считать, равные) $t_0 < t_1 < t_2 < \ldots < t_{n-1} < t_n < \ldots$, на текущем отрезке $(t_{n-1}; t_n]$ для каждого источника $\Re$ измеряются и накапливаются данные о времени выполнения поступающих запросов, по наступлению момента t_n из накопленных данных формируется нужное наблюдение — вычисляется среднее или суммарное время ожидания отклика. Длина интервала $(t_{n-1}; t_n]$ выбирается достаточно большой (5, 10 или даже 30 минут), чтобы затраты вычислительных ресурсов, выделенных портальным приложениям на оптимизацию собственной работы, не оказывали влияния на выполнение прямой функциональности портала.

Задание чисел $C_Б, C_C, C_M$ и $\Sigma_Б, \Sigma_C, \Sigma_M$ само по себе представляет отдельную задачу, обсуждаемую далее. Пока ограничимся пониманием их физического смысла и будем предполагать, что они известны.

Далее нужно сформировать модель для классифицирующей функции $I(\Re \in \theta)$. Эта модель должна описывать закон, по которому источник $\Re$ на текущем интервале наблюдения $(t_{n-1}; t_n]$ относится к тому или иному классу эффективности θ. Как упоминалось выше, предложить эту модель, исходя из физического смысла показателя эффективности, затруднительно, т. е. хотя этот закон очевидно существует, но для постороннего наблюдателя является «черным ящиком». По этой причине поступим так, как рекомендует классическая теория идентификации систем [10], а именно, опишем этот закон простой параметрической моделью. Будем считать, что для источника $\Re$ существует случайный процесс X_t, эволюция которого описывается уравнением авторегрессии первого порядка:

$$X_{t_n} = AX_{t_{n-1}} + Q + BV_t, \tag{2}$$

где A, Q, B — известные неслучайные (заданные или подлежащие последующей идентификации) числа, V_t — случайное возмущение.

Смысл процесса X_t можно прокомментировать следующими соображениями. Под объективной числовой характеристикой эффективности источника $\Re$ в момент t (обозначим ее T_t) можно понимать некоторое абсолютное время, затрачиваемое источником на выполнение запросов на текущем интервале наблюдения $(t_{n-1}; t_n]$. Абсолютность T_t означает возможность влияния на его величину только тех факторов неопределенности, которые связаны непосредственно с источником (семантика выполняемого запроса, полнота данных источника, эксплуатационные характеристики аппаратного обеспечения и т. п.). Если бы величина T_t была известна, то индикаторная функция $I(\Re \in \theta)$ вычислялась бы путем определения, в какой из заданных числовых интервалов эта величина попала. С учетом разнородности источников целесообразным представляется масштабировать показатель T_t. В похожих ситуациях, например в финансовых приложениях [11], традиционно переходят от рассмотрения абсолютного показателя к относительному: $\dfrac{T_{t_n} - T_{t_{n-1}}}{T_{t_n}}$, который в нашем случае можно назвать «относительной эффективностью» или «индексом эффективности». Рассматривая динамику этого индекса, можно считать,

что $X_{t_n} = X_{t_{n-1}} + \frac{T_{t_n} - T_{t_{n-1}}}{T_{t_n}}$. Используя для описания X_t модель (2), остается выбирать параметры, добиваясь адекватности модели наблюдения реально получаемым данным.

Класс, к которому следует отнести источник $\Re$, т. е. функцию классификации $I(\Re \in \theta)$, определим следующим образом. Область значений процесса X_t — числовую прямую R^1 — разобьем на непересекающиеся интервалы, поставив каждому классу эффективности в соответствие один интервал, и значение $I(\Re \in \theta)$ определим как принадлежность X_t соответствующему интервалу. В рассматриваемом примере нужны три интервала $\Delta_Б, \Delta_C, \Delta_M$: $\Delta_Б \cup \Delta_C \cup \Delta_M = (-\infty; +\infty)$, вектор-функция классификации источника имеет вид:

$$\mathbf{\Theta}(X_t) = \begin{pmatrix} I(\Re \in \theta_Б) \\ I(\Re \in \theta_C) \\ I(\Re \in \theta_M) \end{pmatrix} = \begin{pmatrix} I(X_t \in \Delta_Б) \\ I(X_t \in \Delta_C) \\ I(X_t \in \Delta_M) \end{pmatrix}. \qquad (3)$$

Таким образом, рассматриваемая задача оптимального распределения множества запросов по пулу сведена к задаче оценивания индикаторной функции (*3*) *по наблюдениям* (*1*), (*2*).

2.2. Формальная постановка и решение задачи фильтрации — оценивания эффективности источника

Суммируя соображения предыдущего пункта и внося непринципиальные обобщения, получаем следующую постановку задачи фильтрации в дискретной стохастической системе наблюдения.

Обозначим далее:

- $M[X], M[X \mid \Im]$ — соответственно, безусловное математическое ожидание случайной величины X и условное математическое ожидание случайной величины X относительно σ-алгебры $\Im$;
- $P(\cdot), P(\cdot \mid \Im)$ — соответственно, вероятностная и условная относительно σ-алгебры $\Im$ вероятностная меры, заданные на подходящем вероятностном пространстве;
- X^T — операция транспонирования вектора (матрицы) X;
- $col(x^1,...,x^n) \stackrel{\Delta}{=} (x^1,...,x^n)^T$ — вектор-столбец с элементами $x^1,...,x^n$;

- $row(x^1,...,x^n) \stackrel{\Delta}{=} (x^1,...,x^n)$ — вектор-строка с элементами $x^1,...,x^n$;
- $e_k = (0,...,0,1,0,...0)^T$ — единичный вектор в пространстве R^n, все координаты которого равны 0, а k-я равна 1, $1 \le k \le n$.

Уравнения состояния динамической системы наблюдения имеют вид

$$x_t = a_t x_{t-1} + q_t + b_t v_t, \quad t = 1, 2,..., \tag{4}$$

где $\{a_t\}, \{q_t\}, \{b_t\}$ — известные неслучайные последовательности, $\{v_t\}$ — стандартный дискретный белый шум в узком смысле [12], сечения которого имеют плотность вероятности $\varphi_v(\cdot)$, x_0 — случайная величина, не зависящая от $\{v_t\}$ и имеющая плотность вероятности $\psi_0(\cdot)$.

Область значений R^1 процесса x_t разбита на пересекающиеся интервалы

$$\Delta_k: -\infty = \alpha_1 < \alpha_2 < ... < \alpha_n < \alpha_{n+1} = +\infty, \Delta_k = (\alpha_k, \alpha_{k+1}], k = 1, ..., n-1, \Delta_n = (\alpha_n, +\infty),$$

и определена индикаторная функция $\Theta(x)$:

$$\Theta(x) = col(I_{\Delta_1}(x),..., I_{\Delta_n}(x)), \; I_{\Delta_k}(x) = \begin{cases} 1, \text{ если } x \in \Delta_k, \\ 0, \text{ если } x \notin \Delta_k. \end{cases} \tag{5}$$

Уравнения наблюдений имеют вид

$$y_t = c_t \Theta(x_t) + \sigma_t \Theta(x_t) w_t, \tag{6}$$

где $\{c_t\}, \{\sigma_t\}$, $c_t = row(c_t^1,...,c_t^n), \sigma_t = row(\sigma_t^1,...,\sigma_t^n)$ — известные неслучайные последовательности, $\{w_t\}$ — стандартный дискретный белый шум в узком смысле, сечения которого имеют плотность вероятности $\varphi_w(\cdot)$ и не зависят от $\{v_t\}, x_0$.

Принимая во внимание множественный характер возмущений, моделируемых шумами $\{v_t\}$ и $\{w_t\}$, а также независимость отдельных факторов, вносящих вклад в данные шумы, можно, учитывая центральную предельную теорему [12], считать, что $\varphi_v(\cdot), \varphi_w(\cdot)$ — стандартные гауссовские плотности вероятности (данное предположением используется при практической реализации описываемого алгоритма, но на результат данного раздела не влияет).

Обозначим:

- $\Im_t^y \overset{\Delta}{=} \sigma\{y_\tau, \tau \le t\}$ — σ -алгебра, порожденная наблюдениями $y_\tau, \tau \le t$,
- $\bar{\psi}_x(x,t)$ — условная плотность вероятности x_t относительно $\Im_{t-1}^y$,
- $\bar{\psi}_y(y,t)$ — условная плотность вероятности y_t относительно $\Im_{t-1}^y$,
- $\bar{\psi}_z(x,y,t)$ — условная плотность вероятности $z_t = col(x_t, y_t)$ относительно $\Im_{t-1}^y$,
- $\hat{\psi}_x(x,t)$ — условная плотность вероятности x_t относительно $\Im_t^y$.

Отметим, что плотности $\bar{\psi}_x(x,t)$, $\bar{\psi}_y(y,t)$ и $\bar{\psi}_z(x,y,t)$ обеспечивают решение задач прогнозирования в системе (4)–(6), плотность $\hat{\psi}_x(x,t)$ — решение задачи фильтрации, в том числе и оценивания индикаторной функции $\Theta(x)$ по наблюдениям $y_\tau, \tau \le t$.

Теорема 1. Пусть для системы наблюдения (4)–(6) выполнено:

1) $\exists \varepsilon > 0 : b_\tau \ge \varepsilon \ \forall \tau \le t$,

2) $\exists \delta > 0 : \min\limits_{1 \le k \le n} \sigma_\tau^k \ge \delta \ \forall \tau \le t$.

Тогда условная плотность вероятности $\hat{\psi}_x(x,t)$ существует и описывается уравнением

$$\hat{\psi}_x(x,t) = \frac{\sum\limits_{k=1}^{n} \frac{\Theta^T(x) e_k}{\sigma_t^k} \varphi_w\left(\frac{y_t - c_t^k}{\sigma_t^k}\right) \int\limits_{R^1} \hat{\psi}_x(\xi, t-1)\, \varphi_v\left(\frac{x - a_t\xi - q_t}{b_t}\right) d\xi}{\sum\limits_{k=1}^{n} \frac{1}{\sigma_t^k} \varphi_w\left(\frac{y_t - c_t^k}{\sigma_t^k}\right) \int\limits_{\Delta_k} \int\limits_{R^1} \hat{\psi}_x(\xi, t-1)\, \varphi_v\left(\frac{x - a_t\xi - q_t}{b_t}\right) d\xi\, dx}. \qquad (7)$$

Доказательство.

Для условной функции распределения z_t относительно $\Im_{t-1}^y$ имеем:

$$P(x_t \le x,\ y_t \le y \mid \Im_{t-1}^y) = P(a_t x_{t-1} + q_t + b_t v_t \le x,\ c_t\Theta(x_t) + \sigma_t\Theta(x_t) w_t \le y \mid \Im_{t-1}^y) =$$

$$= P\left(v_t \le \frac{x - a_t x_{t-1} - q_t}{b_t},\ w_t \le \frac{y - c_t\Theta(x_t)}{\sigma_t\Theta(x_t)} \mid \Im_{t-1}^y\right) =$$

$$= P\left(v_t \le \frac{x - a_t x_{t-1} - q_t}{b_t}, w_t \le \sum_{k=1}^{n} \frac{y - c_t^k}{\sigma_t^k} I_{\Delta_k}(x_t) \mid \Im_{t-1}^y \right) =$$

$$= \sum_{k=1}^{n} P\left(v_t \le \frac{x - a_t x_{t-1} - q_t}{b_t}, w_t \le \frac{y - c_t^k}{\sigma_t^k}, x_t \in \Delta_k \mid \Im_{t-1}^y \right) =$$

$$= \sum_{k=1}^{n} P\left(v_t \le \frac{x - a_t x_{t-1} - q_t}{b_t}, w_t \le \frac{y - c_t^k}{\sigma_t^k}, \alpha_k < a_t x_{t-1} + q_t + b_t v_t \le \alpha_{k+1} \mid \Im_{t-1}^y \right) =$$

$$= \sum_{k=1}^{n} P\left(\frac{\alpha_k - a_t x_{t-1} - q_t}{b_t} < v_t \le \frac{\min(x, \alpha_{k+1}) - a_t x_{t-1} - q_t}{b_t}, w_t \le \frac{y - c_t^k}{\sigma_t^k} \mid \Im_{t-1}^y \right) =$$

$$= \sum_{k=1}^{n} \int_{R^1} \hat{\psi}_x(\xi, t-1) I(x - \alpha_k) \int_{\frac{\alpha_k - a_t\xi - q_t}{b_t}}^{\frac{\min(x,\alpha_{k+1}) - a_t\xi - q_t}{b_t}} \varphi_v(v) dv \, d\xi \int_{-\infty}^{\frac{y - c_t^k}{\sigma_t^k}} \varphi_w(w) dw, \qquad (8)$$

где учтены предположение о существовании условной плотности вероятности $\hat{\psi}_x(x, t-1)$ и независимость величин v_t и w_t друг от друга и от σ-алгебры $\Im_{t-1}^y$, индикаторная функция $I(x - \alpha_k) = \begin{cases} 1, x > \alpha_k, \\ 0, x \le \alpha_k \end{cases}$ введена для обнуления вероятности в случае $\alpha_k \ge \min(x, \alpha_{k+1})$. Соотношение (8) следует непосредственно из существования и свойств регулярных условных распределений для случайных последовательностей [13].

Из выражения (8), дифференцируя интегралы с переменными верхними пределами по x и по y и учитывая, что в каждом слагаемом $x \in \Delta_k$, получаем, что существует условная плотность вероятности $\bar{\psi}_z(x, y, t)$ следующего вида:

$$\bar{\psi}_z(x, y, t) = \sum_{k=1}^{n} \frac{1}{\sigma_t^k b_t} \int_{R^1} \hat{\psi}_x(\xi, t-1) I_{\Delta_k}(x) \varphi_v\left(\frac{x - a_t\xi - q_t}{b_t} \right) \varphi_w\left(\frac{y - c_t^k}{\sigma_t^k} \right) d\xi. \qquad (9)$$

Учитывая в равенстве (9) введенные обозначения для единичных векторов и проводя очевидные преобразования, получаем:

$$\bar{\psi}_z(x, y, t) = \sum_{k=1}^{n} \frac{\Theta^T(x) e_k}{\sigma_t^k b_t} \varphi_w\left(\frac{y - c_t^k}{\sigma_t^k} \right) \int_{R^1} \hat{\psi}_x(\xi, t-1) \varphi_v\left(\frac{x - a_t\xi - q_t}{b_t} \right) d\xi. \qquad (10)$$

Из (10) с учетом определения (5) путем интегрирования по x получается маргинальная условная плотность $\bar{\psi}_y(y,t)$:

$$\bar{\psi}_y(y,t) = \sum_{k=1}^{n} \frac{1}{\sigma_t^k b_t} \varphi_w\left(\frac{y - c_t^k}{\sigma_t^k}\right) \int\limits_{\Delta_k} \int\limits_{R^1} \hat{\psi}_x(\xi, t-1)\, \varphi_v\left(\frac{x - a_t\xi - q_t}{b_t}\right) d\xi\, dx. \quad (11)$$

Окончательно соотношение (7) получаем по формуле Байеса для плотностей вероятности [12] как отношение (10) и (11), где $y = y_t$. ***Теорема 1 доказана.***

Замечание 1. В рекуррентных соотношениях (7) в качестве начальной плотности вероятности используется $\hat{\psi}_x(x,0) = \psi_0(x)$. Эти соотношения полностью описывают рекуррентную процедуру пересчета искомой условной плотности вероятности x_t относительно $\Im_t^y$ — решения задачи оптимальной фильтрации.

Замечание 2. При отказе от условий 1)–2) в теореме 1 (отделимости от нуля) уравнение (7) не изменится принципиально: полученная формула будет иметь место с учетом того, что интегрирование в (7) проводится не по лебеговой мере, а по некоторой абстрактной (сингулярной) мере. Однако для целей данной работы отделимость от нуля диффузионных коэффициентов, очевидно, необходима.

Следствие 1. Наряду с задачей фильтрации (выражением для $\hat{\psi}_x(x,t)$) решены также задачи прогнозирования. Именно, уравнение (10) позволяет решать задачу прогнозирования z_t по наблюдениям y_τ, $\tau \le t-1$, уравнение (11) — задачу прогнозирования y_t по наблюдениям $y_\tau, \tau \le t-1$, а выражение для $\bar{\psi}_x(x,t)$ легко получить из (10) интегрированием $\bar{\psi}_z(x,y,t)$ по y.

Следствие 2. Оптимальная в среднем квадратическом смысле оценка $\hat{x}_t$ состояния x_t по наблюдениям y_τ, $\tau \le t$ (условное математическое ожидание x_t относительно $\Im_t^y$ [12]) имеет вид

$$\hat{x}_t = \int\limits_{R^1} x\hat{\psi}_x(x,t)\, dx. \quad (12)$$

Следствие 3. Оптимальная в среднем квадратическом смысле оценка индикатора принадлежности источника $\Re$ к k-му классу «эффективности» по наблюдениям y_τ, $\tau \le t$, $\hat{\Theta}_k(t) = M[\Theta^T(x_t)e_k \mid \Im_t^y] = P(x_t \in \Delta_k \mid \Im_t^y)$ имеет вид

$$\hat{\Theta}_k(t) = \int\limits_{\Delta_k} \hat{\psi}_x(x,t)\,dx. \qquad (13)$$

После вычисления оценок $\hat{\Theta}_k(t)$ по формуле (13) источник $\Re$ в момент t относится к классу $\hat{k}(t) = \arg\max\limits_{1\le k\le n} \hat{\Theta}_k(t)$.

2.3. Определение параметров в задаче оценивания эффективности источника

Задача фильтрации в предыдущем пункте решена в предположении, что числовые последовательности $\{a_t\}, \{q_t\}, \{b_t\}, \{c_t\}, \{\sigma_t\}$ известны. Действительно, имеется принципиальная возможность определить эти параметры путем предварительного анализа среды, в которой работает портал, и формулировки некоторых выводов, позволяющих сформировать априорную оценку искомых величин. Соответствующую процедуру экспертного оценивания возможно реализовать на основе следующих соображений (в терминах предложенного выше примера).

Заметим, во-первых, что значения величины X_{t_n} из (2) могут быть выбраны достаточно произвольно, поскольку непосредственно эти значения не используются, а нужны только для принятия решения об отношении источника $\Re$ к определенному классу эффективности. Это значит, что и интервалы $\Delta_Б, \Delta_C, \Delta_M$ могут быть выбраны достаточно произвольно. Например, положим $\Delta_Б = (-\infty, -10]$, $\Delta_C = (-10, 10]$, $\Delta_M = (10, +\infty]$. Добиться надо лишь адекватности уравнения наблюдения (1) реально наблюдаемым данным. Будем считать, что по закону (1) наблюдается среднее время выполнения запроса источником $\Re$ в текущий момент времени. Предполагая, что шум W_t имеет единичную дисперсию и нулевое математическое ожидание, числа $C_Б, C_C, C_M$ и $\Sigma_Б, \Sigma_C, \Sigma_M$ выберем следующим образом. $C_Б, C_C, C_M$ — среднее время выполнения запроса источником, отнесенным, соответственно, к классу $\theta_Б, \theta_C, \theta_M$ в некотором стационарном режиме функционирования портала. $\Sigma_Б, \Sigma_C, \Sigma_M$ — среднее отклонение, характерное для источника из соответствующего класса в том же стационарном режиме. Эти величины эксперт может определить, собрав соответствующую статистику в реально работающей системе, либо смоделировать нужный результат, подготовив блок «типовых» пользовательских запросов.

Теперь остается сформировать уравнение состояния (2) для X_{t_n} так, чтобы распределение его значений по интервалам $\Delta_Б, \Delta_C, \Delta_M$ соответство-

вало реальным частотам пребывания источника в соответствующих классах. Шум V_t также будем предполагать стандартным, т. е. имеющим нулевое математическое ожидание и единичную дисперсию. Параметр A выберем так, чтобы обеспечить, прежде всего, устойчивость авторегрессии, что должно имитировать предположение о существовании стационарного режима функционирования портала. В этом предположении не должно изменяться среднее значение процесса X_{t_n}, т. е. $M[X_{t_n}] = M[X_{t_{n-1}}] = Q/(1-A)$. Аналогично постоянной должна быть и дисперсия X_{t_n}, т. е. $D[X_{t_n}] = D[X_{t_{n-1}}] = B^2/(1-A^2)$. Определение чисел A, Q, B также должно основываться на экспертном представлении о вероятности отнесения источника $\Re$ к тому или иному классу при смене внешних условий функционирования. С учетом наличия стационарного распределения у авторегрессии и выбранных интервалов $\Delta_Б, \Delta_C, \Delta_M$ числа A, Q, B нужно подобрать так, чтобы обеспечить равенство вычисленной для стационарного распределения $P(X_{t_n} \in \Delta_k)$ определенной экспертом вероятности отнесения источника $\Re$ к соответствующему классу θ_k. Кроме того, значение параметра A позволяет задать характер устойчивости авторегрессии (в некотором смысле, скорость, с которой источник $\Re$, покинув один класс, вернется обратно). Так, например, если источник $\Re$ предполагается отнести к классу θ_C со средней характеристикой эффективности $M[X_{t_n}] = 1,$ то характер устойчивости авторегрессии можно определить, например, задав $A = 0,9,$ т. е. достаточно близким к 1, чтобы моделировать длительное время пребывания источника в каждом из классов эффективности. При этом следует, очевидно, выбрать $Q = 0,1.$ Дисперсию $D[X_{t_n}] \approx 5B^2$ будем задавать, отображая экспертное представление о вероятности отнесения источника $\Re$ к соответствующим классам эффективности (с учетом выбранных интервалов). Если такая вероятность велика, то можно положить $B = 2, 3, 4, ...,$ если мала, то можно положить $B = 1; 0,7; 0,5; ...$ Для другого источника можно положить $A = 0,1$, чтобы моделировать кратковременное пребывание источника в каждом из классов эффективности, и аналогично подобрать Q, B.

По описанной процедуре определения параметров можно сделать следующие замечания:

- предложенная схема экспертного определения параметров довольно трудоемка, но вполне реализуема;
- очевидным принципиальным соображением, в любом случае пригодным для практического применения, является замечание о возможно-

сти произвольного выбора интервалов $\Delta_E, \Delta_C, \Delta_M$ (в общем случае $\Delta_k = (\alpha_k, \alpha_{k+1}], k = 1,...,n$);

- рекомендации по экспертному определению параметров модели могут использоваться частично (т. е. только для части параметров модели) в том случае, если будет предложена процедура, обеспечивающая возможность вычисления (идентификации) не заданных экспертом величин.

Таким образом, с одной стороны, установлена возможность обоснованного выбора параметров рассматриваемой модели за счет простейшего статистического анализа среды функционирования портала. С другой стороны, этот анализ может сочетаться вместе с любым методом идентификации. В качестве такого метода с учетом полученного выражения для условной плотности вероятности может использоваться метод максимального правдоподобия [10]. Более предпочтительной представляется постановка задачи байесовской идентификации параметров, ее детальное рассмотрение может стать предметом дальнейших исследований.

3. Оценка активности пользователей портала

3.1. Модель формирования оценки активности пользователей

Данный раздел посвящен второй части общей задачи управления пулом запросов компонента выполнения, а именно, определению числа $L_{HП}$ нитей, выделяемых на выполнение текущей команды, из общего числа $L_П$ нитей пула (см. рис. 1). Зная число X_t пользователей, работающих в момент времени t с порталом, естественным было бы определить $L_{HП} = [L_П / X_t]$, где $[X]$ — целая часть X. Те же, что и выше, рассуждения о целесообразности разбиения интервала работы портала $[t_0; +\infty)$ на отрезки $(t_{n-1}; t_n]$ можно учесть, понимая под X_{t_n} либо число пользователей, активных в момент t_n, либо среднее число пользователей, обращавшихся с командами к порталу за время $(t_{n-1}; t_n]$. С учетом этого обстоятельства соотношение для $L_{HП}$ естественно скорректировать, положив

$$L_{HП} = \max\left(1, \min\left(L_П, \left[\frac{L_П}{X_{t_n}} \frac{t_n - t_{n-1}}{T_{ср}}\right]\right)\right), \qquad (14)$$

где $T_{ср}$ — среднее время выполнения команды на интервале наблюдения $(t_{n-1}; t_n]$, которое измеряется наряду с другими характеристиками функ-

ционирования портала. Метод управления (14), таким образом, обеспечивает некоторый обоснованный выбор числа нитей $L_{НП}$, если текущие параметры функционирования портала допускают такую возможность, или устанавливает для каждого пользователя возможность использования только одной нити в случае существенно увеличенной нагрузки. В последнем «наихудшем» случае все пользователи окажутся в равных условиях, ожидая выполнения команды ровно столько времени, сколько требует последовательное выполнение запросов к источникам, сформированным из заданной пользователем команды.

Проблема, однако, состоит в том, что для использования в (14) непосредственно определить X_{t_n} невозможно, так как работа портала ведется без установления сессионных соединений (stateless). Кроме того, не всегда удается ассоциировать команду с конкретным пользователем: например, вариант портала с полностью открытым доступом на чтение данных принципиально не позволяет отделить ситуацию, когда за интервал наблюдения выполнено десять запросов одним пользователем, от ситуации, когда за это же время выполнено по одному запросу десятью пользователями. Так не происходит, если конкретный пользователь, работая с порталом, направляет команды последовательно: каждая следующая формируется после получения ответа на предыдущую. Однако способа заставить пользователя дожидаться ответа, прежде чем посылать очередную команду, не предусматривается самой технологией: ничто не мешает направить практически одновременно любое количество команд, например, открыв и заполнив сразу несколько поисковых форм.

Таким образом, вместо X_{t_n} приходится использовать оценку $\hat{X}_{t_n}$, формируемую по результатам косвенных наблюдений. В качестве последних использовать можно лишь общее число команд Y_{t_n}, выполненных порталом за время $(t_{n-1}; t_n]$.

Эволюцию X_{t_n} допустимо было бы описывать уравнением авторегрессии первого порядка:

$$X_{t_n} = AX_{t_{n-1}} + Q + BV_{t_n}, \tag{15}$$

предполагая, что среднее число пользователей (определяется коэффициентами A, Q) постоянно, а возмущение V_t моделирует возможные отклонения от этого среднего. Такая модель не может быть признана адекватной из-за хорошо известного администраторам интернет-ресурсов факта спонтанного изменения активности пользователей. Именно, в отдельные промежутки времени интерес пользователей (активность) к тому или иному сайту меняется скачкообразно, остается более-менее постоянным некото-

рое время, а затем изменяется вновь. При этом можно считать, что на каждом таком интервале активность в некотором смысле постоянна (изменяется сравнительно мало).

Для моделирования этого обстоятельства, как и выше, воспользуемся идеей классификации рассматриваемого показателя X_{t_n}. Именно, будем считать, что есть несколько (например, три) режима пользовательской активности:

- $\theta_{<10}$ — режим слабой активности (например, менее 10 пользователей);
- θ_{100} — режим повседневной активности (описывается средним характерным числом пользователей, регулярно использующих портал в рабочее время);
- $\theta_{>100}$ — режим повышенной активности.

В предположении, что в конкретном режиме число пользователей X_{t_n} описывается уравнением авторегрессии первого порядка, получаем вместо (15) эволюционную модель на основе индикаторов:

$$X_{t_n} = \mathbf{A} \cdot \begin{pmatrix} I(\textit{режим}\ \theta_{<10}) \\ I(\textit{режим}\ \theta_{100}) \\ I(\textit{режим}\ \theta_{>100}) \end{pmatrix} \cdot X_{t_{n-1}} + \mathbf{Q} \cdot \begin{pmatrix} I(\textit{режим}\ \theta_{<10}) \\ I(\textit{режим}\ \theta_{100}) \\ I(\textit{режим}\ \theta_{>100}) \end{pmatrix} + \mathbf{B} \cdot \begin{pmatrix} I(\textit{режим}\ \theta_{<10}) \\ I(\textit{режим}\ \theta_{100}) \\ I(\textit{режим}\ \theta_{>100}) \end{pmatrix} \cdot V_{t_n}, \tag{16}$$

где вектор-строки $\mathbf{A} = (A_{<10}, A_{100}, A_{>100})$ и $\mathbf{Q} = (Q_{<10}, Q_{100}, Q_{>100})$ определяют среднее число пользователей, характерное для соответствующего режима, V_t — возмущение (шум), моделирующее изменения уровня пользовательской активности, $\mathbf{B} = (B_{<10}, B_{100}, B_{>100})$ — вектор-строка, элементы которой определяют уровень отклонения фактического числа пользователей в зависимости от текущего режима. Индикаторные функции $I(\textit{режим}\ \theta)$ в (16) принимают значение 1, если текущий режим активности θ, и 0 противном случае.

Задать индикаторную функцию $I(\textit{режим}\ \theta)$ естественно будет путем разбиения области значений процесса X_t на непересекающиеся интервалы и установления соответствия каждому режиму активности одного интервала. В рассматриваемом примере выбраны характерные для выделения трех режимов числа пользователей 10 и 100, поэтому три интервала $\Delta_{<10}, \Delta_{100}, \Delta_{>100}$ можно определить так: $\Delta_{<10} = (-\infty, 10]$, $\Delta_{100} = (10, 100]$, $\Delta_{100} = (100, +\infty]$. Тогда вектор-функция определения режима активности имеет вид:

$$\mathbf{\Theta}(X_t)=\begin{pmatrix} I(\text{режим } \theta_{<10}) \\ I(\text{режим } \theta_{100}) \\ I(\text{режим } \theta_{>100}) \end{pmatrix}=\begin{pmatrix} I(X_t\in\Delta_{<10}) \\ I(X_t\in\Delta_{100}) \\ I(X_t\in\Delta_{>100}) \end{pmatrix}. \qquad (17)$$

Для модели наблюдения за состоянием (16) можно воспользоваться представляющимся достаточно обоснованным предположением о линейной зависимости между числом активных пользователей X_{t_n} и числом команд Y_{t_n}, выполненных порталом за интервал наблюдения:

$$Y_{t_n}=CX_{t_n}+\Sigma W_{t_n}, \qquad (18)$$

где параметр C определяет среднее число команд, формируемых одним пользователем за интервал наблюдения $(t_{n-1};t_n]$, W_t — возмущение (шум), моделирующее отклонения числа команд от заданного среднего уровня, Σ — среднее квадратическое отклонение этого возмущения.

Таким образом, рассматриваемая задача оптимального определения числа нитей (*части общего пула*)*, отводимых под выполнение одной пользовательской команды, сведена к задаче оценивания состояния случайного процесса* X_{t_n} (*16*) *по наблюдениям* (*18*)*.*

3.2. Формальная постановка и решение задачи фильтрации — оценивания активности пользователей

Суммируя соображения предыдущего пункта и внося непринципиальные обобщения, получаем следующую постановку задачи фильтрации в дискретной стохастической системе наблюдения (будем также использовать введенные выше обозначения для вероятностных характеристик).

Уравнения состояния динамической системы наблюдения имеют вид

$$x_t=a_t\Theta(x_{t-1})x_{t-1}+q_t\Theta(x_{t-1})+b_t\Theta(x_{t-1})v_t,\ t=1,2,..., \qquad (19)$$

где $\{a_t\},\{q_t\},\{b_t\}$, $a_t=row(a_t^1,...,a_t^n)$, $q_t=row(q_t^1,...,q_t^n)$, $b_t=row(b_t^1,...,b_t^n)$ — известные неслучайные последовательности, $\{v_t\}$ — стандартный дискретный белый шум в узком смысле, сечения которого имеют плотность вероятности $\varphi_v(\cdot)$, x_0 — случайная величина, не зависящая от $\{v_t\}$ и имеющая плотность вероятности $\psi_0(\cdot)$, $\Theta(x)$ определена в (5).

Уравнения наблюдений имеют вид

$$y_t=c_tx_t+\sigma_tw_t, \qquad (20)$$

где $\{c_t\}, \{\sigma_t\}$ — известные неслучайные последовательности, $\{w_t\}$ — стандартный дискретный белый шум в узком смысле, сечения которого имеют плотность вероятности $\varphi_w(\cdot)$ и не зависят от $\{v_t\}$, x_0.

Теорема 2. Пусть для системы наблюдения (19), (20) выполнено:

1) $\exists \varepsilon > 0 : \min\limits_{1 \le k \le n} b_\tau^k \ge \varepsilon \ \forall \tau \le t,$

2) $\exists \delta > 0 : \sigma_\tau \ge \delta \ \forall \tau \le t.$

Тогда условная плотность вероятности $\hat{\psi}_x(x,t)$ существует и описывается уравнением

$$\hat{\psi}_x(x,t) = \frac{\sum\limits_{k=1}^{n} \frac{1}{b_t^k} \varphi_w\left(\frac{y_t - c_t x}{\sigma_t}\right) \int\limits_{\Delta_k} \hat{\psi}_x(\xi, t-1)\, \varphi_v\left(\frac{x - a_t^k \xi - q_t^k}{b_t^k}\right) d\xi}{\sum\limits_{k=1}^{n} \frac{1}{b_t^k} \int\limits_{R^1} \varphi_w\left(\frac{y_t - c_t x}{\sigma_t}\right) \int\limits_{\Delta_k} \hat{\psi}_x(\xi, t-1)\, \varphi_v\left(\frac{x - a_t^k \xi - q_t^k}{b_t^k}\right) d\xi\, dx} \qquad (21)$$

Доказательство.

Для условной функции распределения z_t относительно $\Im_{t-1}^y$ имеем:

$$\begin{aligned}
&P(x_t \le x, y_t \le y \mid \Im_{t-1}^y) = \\
&= P(a_t \Theta(x_{t-1}) x_{t-1} + q_t \Theta(x_{t-1}) + b_t \Theta(x_{t-1}) v_t \le x, c_t x_t + \sigma_t w_t \le y \mid \Im_{t-1}^y) = \\
&= P\left(v_t \le \frac{x - a_t \Theta(x_{t-1}) - q_t \Theta(x_{t-1})}{b_t \Theta(x_{t-1})}, w_t \le \frac{y - c_t x_t}{\sigma_t} \mid \Im_{t-1}^y\right) = \\
&= P\left(v_t \le \sum_{k=1}^{n} \frac{x - a_t^k x_{t-1} - q_t^k}{b_t^k} I_{\Delta_k}(x_{t-1}), w_t \le \frac{y - c_t x_t}{\sigma_t} \mid \Im_{t-1}^y\right) = \\
&= \sum_{k=1}^{n} P\left(v_t \le \frac{x - a_t^k x_{t-1} - q_t^k}{b_t^k}, w_t \le \frac{y - c_t x_t}{\sigma_t}, x_{t-1} \in \Delta_k \mid \Im_{t-1}^y\right) = \\
&= \sum_{k=1}^{n} P\left(v_t \le \frac{x - a_t^k x_{t-1} - q_t^k}{b_t^k}, w_t \le \frac{y - c_t (a_t^k x_{t-1} + q_t^k + b_t^k v_t)}{\sigma_t}, x_{t-1} \in \Delta_k \mid \Im_{t-1}^y\right) = \\
&= \sum_{k=1}^{n} \int\limits_{\Delta_k} \hat{\psi}_x(\xi, t-1) \int\limits_{-\infty}^{\frac{x - a_t^k \xi - q_t^k}{b_t^k}} \varphi_v(v) \int\limits_{-\infty}^{\frac{y - c_t (a_t^k \xi + q_t^k + b_t^k v)}{\sigma_t}} \varphi_w(w)\, dw\, dv\, d\xi, \qquad (22)
\end{aligned}$$

где учтены предположение о существовании условной плотности вероятности $\hat{\psi}_x(x,t-1)$ и независимость величин v_t и w_t друг от друга и от σ-алгебры $\Im_{t-1}^y$. Соотношение (22) следует непосредственно из существования и свойств регулярных условных распределений для случайных последовательностей [13].

Из выражения (22), дифференцируя интегралы с переменными верхними пределами по x и y, получаем, что существует условная плотность вероятности $\bar{\psi}_z(x,y,t)$ следующего вида:

$$\bar{\psi}_z(x,y,t) =$$

$$= \sum_{k=1}^{n} \frac{1}{\sigma_t b_t^k} \int_{\Delta_k} \hat{\psi}_x(\xi,t-1)\varphi_v\left(\frac{x-a_t^k\xi-q_t^k}{b_t^k}\right)\varphi_w\left(\frac{y-c_t\left(a_t^k\xi+q_t^k+b_t^k\dfrac{x-a_t^k\xi-q_t^k}{b_t^k}\right)}{\sigma_t}\right)d\xi, \quad (23)$$

или, после упрощения:

$$\bar{\psi}_z(x,y,t) = \sum_{k=1}^{n} \frac{1}{\sigma_t b_t^k} \varphi_w\left(\frac{y-c_t x}{\sigma_t}\right) \int_{\Delta_k} \hat{\psi}_x(\xi,t-1)\varphi_v\left(\frac{x-a_t^k\xi-q_t^k}{b_t^k}\right)d\xi. \quad (24)$$

Из (24) путем интегрирования по x получается маргинальная условная плотность $\bar{\psi}_y(y,t)$:

$$\bar{\psi}_y(y,t) = \sum_{k=1}^{n} \frac{1}{\sigma_t b_t^k} \int_{R^1} \varphi_w\left(\frac{y-c_t x}{\sigma_t}\right) \int_{\Delta_k} \hat{\psi}_x(\xi,t-1)\varphi_v\left(\frac{x-a_t^k\xi-q_t^k}{b_t^k}\right)d\xi\, dx. \quad (25)$$

Окончательно соотношение (21) получаем по формуле Байеса для плотностей вероятности [12] как отношение (24) и (25), где $y = y_t$. ***Теорема 2 доказана.***

Отметим, что требующаяся для реализации управляющего воздействия оценка $\hat{x}_t$ состояния x_t по наблюдениям $y_\tau, \tau \le t$ имеет тот же вид (12) из следствия 2.

***Следствие 4* (*гауссовский случай*).** Практически большой интерес представляет система наблюдения (19), (20) с гауссовскими возмущениями $\{v_t\}, \{w_t\}$. Уравнения оптимальной фильтрации (21) и (12) состояния x_t по наблюдениям $y_\tau, \tau \le t$ в этом случае решаются с учетом того, что

функции $\varphi_v(\cdot)$ и $\varphi_w(\cdot)$ являются плотностями вероятности стандартного нормального распределения, благодаря чему (21) и (12) могут быть существенно упрощены. В знаменателе (21) поменяем порядок интегрирования и представим $\overline{\psi}_y(y,t)$ в следующем виде:

$$\overline{\psi}_y(y,t)=\sum_{k=1}^{n}\int\limits_{\Delta_k}\hat{\psi}_x(\xi,t-1)\left(\frac{1}{\sigma_t b_t^k}\int\limits_{R^1}\varphi_w\left(\frac{y-c_t x}{\sigma_t}\right)\varphi_v\left(\frac{x-a_t^k\xi-q_t^k}{b_t^k}\right)dx\right)d\xi \quad (26)$$

Для интегрирования в (12) рассмотрим отдельно числитель и также поменяем порядок интегрирования:

$$\hat{x}_t=\frac{\sum\limits_{k=1}^{n}\int\limits_{\Delta_k}\hat{\psi}_x(\xi,t-1)\frac{1}{\sigma_t b_t^k}\left(\int\limits_{R^1}x\varphi_w\left(\frac{y_t-c_t x}{\sigma_t}\right)\varphi_v\left(\frac{x-a_t^k\xi-q_t^k}{b_t^k}\right)dx\right)d\xi}{\overline{\psi}_y(y_t,t)} \quad (27)$$

Вычислим в (26), (27) внутренние интегралы. Для этого рассмотрим две независимые случайные величины V и W, имеющие стандартное гауссовское распределение с плотностями вероятности $\varphi_v(x)$ и $\varphi_w(y)$ соответственно. Также определим случайные величины $X\overset{\Delta}{=}b_t^k V+a_t^k\xi+q_t^k$, $Y\overset{\Delta}{=}\sigma_t W+c_t X$ и обозначим:

- $f_X(x)$ — плотность вероятности X;
- $f_Y(y)$ — плотность вероятности Y;
- $f_{X,Y}(x,y)$ — совместная плотность вероятности X и Y;
- $f_{X|Y}(x\,|\,y)$ —условная плотность вероятности X относительно Y;
- $f_{Y|X}(y\,|\,x)$ — условная плотность вероятности Y относительно X.

Для вычисления внутреннего интеграла в (26) запишем

$$f_{X,Y}(x,y)=f_X(x)f_{Y|X}(y\,|\,x)$$

и заметим, что

$$f_X(x)=\frac{1}{b_t^k}\varphi_v\left(\frac{x-a_t^k\xi-q_t^k}{b_t^k}\right) \text{ и } f_{Y|X}(y\,|\,x)=\frac{1}{\sigma_t}\varphi_w\left(\frac{y-c_t x}{\sigma_t}\right).$$

Поскольку

$$\int_{R^1} f_{X,Y}(x,y)dx = f_Y(y) = \frac{1}{\sqrt{\sigma_t^2 + c_t^2\left(b_t^k\right)^2}}\varphi_w\left(\frac{y - c_t(a_t^k\xi - q_t^k)}{\sqrt{\sigma_t^2 + c_t^2\left(b_t^k\right)^2}}\right),$$

то

$$\frac{1}{\sigma_t b_t^k}\int_{R^1}\varphi_w\left(\frac{y - c_t x}{\sigma_t}\right)\varphi_v\left(\frac{x - a_t^k\xi - q_t^k}{b_t^k}\right)dx = \frac{1}{\sqrt{\sigma_t^2 + c_t^2\left(b_t^k\right)^2}}\varphi_w\left(\frac{y - c_t(a_t^k\xi - q_t^k)}{\sqrt{\sigma_t^2 + c_t^2\left(b_t^k\right)^2}}\right).$$

Таким образом, для $\bar{\psi}_y(y,t)$ из (26) получаем:

$$\bar{\psi}_y(y,t) = \sum_{k=1}^{n}\frac{1}{\sqrt{\sigma_t^2 + c_t^2\left(b_t^k\right)^2}}\int_{\Delta_k}\hat{\psi}_x(\xi,t-1)\,\varphi_w\left(\frac{y - c_t(a_t^k\xi - q_t^k)}{\sqrt{\sigma_t^2 + c_t^2\left(b_t^k\right)^2}}\right)d\xi \qquad (28)$$

Для вычисления внутреннего интеграла в (27) запишем

$$f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x\,|\,y).$$

Тогда

$$\int_{R^1} x f_{X,Y}(x,y)dx = f_Y(y)M[X\,|\,Y = y].$$

Условное математическое ожидание $M[X\,|\,Y = y]$ определяется по теореме о нормальной корреляции [12]. Именно, с учетом записанных выше распределений $f_X(x)$ и $f_Y(y)$, а также очевидного выражения для ковариации X и Y: $M[(X - M[X])(Y - M[Y])] = c_t\left(b_t^k\right)^2$, получаем:

$$M[X\,|\,Y = y] = a_t^k\xi + q_t^k + \frac{c_t\left(b_t^k\right)^2}{\sigma_t^2 + c_t^2\left(b_t^k\right)^2}(y - c_t(a_t^k\xi + q_t^k)).$$

Окончательно для $\hat{x}_t$ из (27) имеем:

$$\hat{x}_t = \frac{\sum_{k=1}^{n} \frac{1}{\sqrt{\sigma_t^2 + c_t^2 \left(b_t^k\right)^2}} \int_{\Delta_k} \hat{\psi}_x(\xi, t-1) \left(\frac{\sigma_t^2 (a_t^k \xi + q_t^k) + c_t \left(b_t^k\right)^2 y_t}{\sigma_t^2 + c_t^2 \left(b_t^k\right)^2} \right) \varphi_w \left(\frac{y_t - c_t (a_t^k \xi - q_t^k)}{\sqrt{\sigma_t^2 + c_t^2 \left(b_t^k\right)^2}} \right) d\xi}{\bar{\psi}_y(y_t, t)} \quad (29)$$

Наконец, вместо (21) остается записать (30), что и завершит рассмотрение гауссовского случая:

$$\hat{\psi}_x(x,t) = \frac{\sum_{k=1}^{n} \frac{1}{\sigma_t b_t^k} \varphi_w \left(\frac{y_t - c_t x}{\sigma_t} \right) \int_{\Delta_k} \hat{\psi}_x(\xi, t-1) \, \varphi_v \left(\frac{x - a_t^k \xi - q_t^k}{b_t^k} \right) d\xi}{\sum_{k=1}^{n} \frac{1}{\sqrt{\sigma_t^2 + c_t^2 \left(b_t^k\right)^2}} \int_{\Delta_k} \hat{\psi}_x(\xi, t-1) \, \varphi_w \left(\frac{y_t - c_t (a_t^k \xi - q_t^k)}{\sqrt{\sigma_t^2 + c_t^2 \left(b_t^k\right)^2}} \right) d\xi} \quad (30)$$

Отметим в заключение, что как и в задаче оценивания эффективности источников в данном случае также возможным является определение параметров рассматриваемой динамической системы путем простейшего статистического анализа портального окружения. Вместе с тем и здесь интерес возможной исследовательской перспективы составляет разработка подхода к идентификации этих параметров.

Заключение

В работе предложено комплексное решение задачи оптимизации функционирования компонента выполнения подсистемы интеграции и поиска Информационного веб-портала — ключевого компонента «интеграционного» ядра программного решения. Предложенные алгоритмы основаны на описании отдельных процессов функционирования веб-портала в форме динамических стохастических систем наблюдения и решении для этих систем задачи фильтрации состояния по косвенным наблюдениям. Показатели функционирования — эффективность «внешних» информационных источников, взаимодействующих с порталом, и характеристика активности пользователей — составляют важную часть описания среды функционирования портала, но не исчерпывают ее полностью. Полученные алгоритмы использованы при выполнении конкретных проектов внедрения Информаци-

онного веб-портала, показавших их практическую пригодность. Вместе с тем анализ результатов, в том числе и соображения, высказанные в данной работе, позволяют видеть очевидные перспективы совершенствования предложенной методики моделирования и оптимизации.

Литература

1. *Оганесян А.* Модели и инструменты интеграции // Открытые системы. 2002. № 11. http://www.osp.ru/os/2002/11/042.htm
2. *Вавилов К.*, *Щербина С.* Web-интеграция // Открытые системы. 2001. № 1. http://www.osp.ru/os/2001/01/178410
3. Информационный Web-портал. Свидетельство об официальной регистрации программы для ЭВМ № 2005612992. Зарегистрировано в Реестре программ для ЭВМ 18.11.2005 г.
4. *Босов А. В.*, *Иванов А. В.* Программная инфраструктура Информационного web-портала РАН // Информатика и ее применения. 2007. Вып. 2. Т. 1. С. 39–53.
5. *Босов А. В.*, *Иванов А. В.* О реализации системы управления содержанием Информационного Web-портала // Информационные технологии и вычислительные системы. М.: ИМВС РАН, 2004. № 4. С. 85–103.
6. *Босов А. В.*, *Полухин А. Н.* О реализации сервиса аутентификации web-портала // Информационные технологии и вычислительные системы. М.: ИМВС РАН, 2005. № 3. С. 50–60.
7. *Босов А. В.*, *Чавтараев Р. Б.* Технология доступа к данным в Информационном web-портале // Информационные технологии и вычислительные системы. М.: ИМВС РАН, 2007. № 1. С. 35–48.
8. *Босов А. В.*, *Чавтараев Р. Б.* Технология поиска данных в информационных источниках web-портала // Информационные технологии и вычислительные системы. М.: ИМВС РАН, 2008. № 1. С. 61–81.
9. *Пугачев В. С.*, *Синицын И. Н.* Теория стохастических систем // М.: Логос, 1971.
10. *Льюнг Л.* Идентификация систем. Теория для пользователя // М.: Наука, 1991.
11. *Ширяев А. Н.* Основы стохастической финансовой математики. Т. 1. Факты. Модели. М.: ФАЗИС, 1998.
12. *Ширяев А. Н.* Вероятность. М.: Наука, 1989.
13. *Гихман И. И.*, *Скороход А. В.* Теория случайных процессов. М.: Наука, 1971. Т. 1.