МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра системного анализа и
автоматического управления

ИССЛЕДОВАНИЕ ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ С УПРАВЛЕНИЕМ ЧИСЛОМ ПОИСКОВЫХ РОБОТОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 481 группы
направления 27.03.03 — Системный анализ и управление
факультета КНиИТ
Исаева Алексея Юрьевича

Научный руководитель
доцент, к. ф.-м. н. _____ Е. С. Рогачко

Заведующий кафедрой
к. ф.-м. н., доцент _____ И. Е. Тананко

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы. Поисковая система - это компьютерная система, предназначенная для поиска информации, предоставляющая пользователю возможность быстрого доступа к ней при помощи поиска в обширной коллекции доступных данных [1]. Одно из наиболее известных применений поисковых систем — веб-сервисы для поиска текстовой или графической информации во Всемирной паутине.

Для поиска информации с помощью поисковой системы пользователь формулирует поисковый запрос [2]. Таким образом, работа поисковой системы заключается в том, чтобы по запросу пользователя найти документы, содержащие либо указанные ключевые слова, либо слова, как-либо связанные с ключевыми словами [3]. При этом поисковая система генерирует страницу результатов поиска. Такая поисковая выдача может содержать различные типы результатов, например: веб-страницы, изображения, аудиофайлы. Некоторые поисковые системы также извлекают информацию из подходящих баз данных и каталогов ресурсов в Интернете.

Для поиска нужных сведений удобнее всего воспользоваться современными поисковыми машинами, которые позволяют быстро обнаружить необходимые сведения и обеспечивают точность и полноту поиска. При работе с этими машинами достаточно задать ключевые слова, наиболее точно отражающие искомую информацию, или составить более сложный запрос из ключевых слов для уточнения области поиска. После ввода запроса на поиск пользователь получает список ссылок на документы в Интернете, в которых содержатся указанные ключевые слова. Обычно ссылки дополняются фрагментами текста из обнаруженного документа, которые часто помогают сразу определить тематику найденной страницы.

В частности, с помощью поисковой системы можно найти много информации о структуре самой поисковой системы и управлении (в том числе распределенном управлении) поисковыми роботами, методах анализа поисковых систем с планированием работы поисковых роботов. В работе [4] предлагается модель устаревания веб-страниц и изучается задача планирования работы поисковых роботов так, чтобы минимизировать устаревание базы данных поисковой системы. Поисковая система тем лучше, чем больше документов, релевантных запросу пользователя, она будет возвращать. Результаты поиска

могут становиться менее релевантными из-за особенностей алгоритмов или вследствие человеческого фактора.

По методам поиска и обслуживания разделяют четыре типа поисковых систем: системы, использующие поисковых роботов, системы, управляемые человеком, гибридные системы и мета-системы.

Цель бакалаврской работы — изучение математической модели информационно-поисковой системы, использующей поисковых роботов, метода анализа системы и метода оптимизации числа поисковых роботов, а также метода динамического управления числом поисковых роботов.

Поставленная цель определила **следующие задачи**:

1. Проанализировать структуру информационно-поисковой системы;
2. Смоделировать информационно-поисковую систему с помощью системы массового обслуживания $M/M/1/K$;
3. Смоделировать информационно-поисковую систему с помощью системы массового обслуживания $M/G/1/K$;
4. Разработать программу для анализа, оптимизации, динамического управления информационно-поисковой системы.

Методологические основы исследования информационно-поисковых систем представлены в работах H. Chu, M. Rosenthal [2], J. Talim, Z. Liu, P. Nain, E. G. Jr. Coffman [4–6].

Практическая значимость бакалаврской работы. Разработанные алгоритмы и программа для анализа модельных систем массового обслуживания позволяют решать задачи анализа, оптимизации, динамического управления для информационно-поисковой системы.

Структура и объем работы. Бакалаврская работа состоит из введения, 4 разделов, заключения, списка использованных источников и приложения. Общий объем работы — 60 страниц, из них 51 страница — основное содержание, включая 18 рисунков, список использованных источников информации — 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Анализ и оптимизация информационно-поисковой системы» посвящен описанию информационно-поисковой системы и моделированию рассматриваемой системы с помощью систем массового обслуживания.

В подразделе 1.1 приведено описание основных компонентов информационно-поисковой системы и описание системы в целом, рассмотрен механизм работы информационно-поисковой системы. Основное внимание уделяется изучению работы поисковых роботов, связанной с актуализацией информации путем индексации новых веб-страниц и переиндексацией измененных/обновленных веб-страниц. Решается задача определения оптимального количества поисковых роботов, удовлетворяющего двум противоположным требованиям: низкого сетевого трафика и актуальности баз данных. На рисунке 1 показана структура рассматриваемой поисковой системы, основными компонентами которой являются механизм индексации и несколько одинаковых поисковых роботов, функционирующих независимо друг от друга, а также представлена математическая модель поисковой системы – одноприборная система массового обслуживания с конечным числом мест для ожидания требований в системе и несколькими источниками поступления требований.

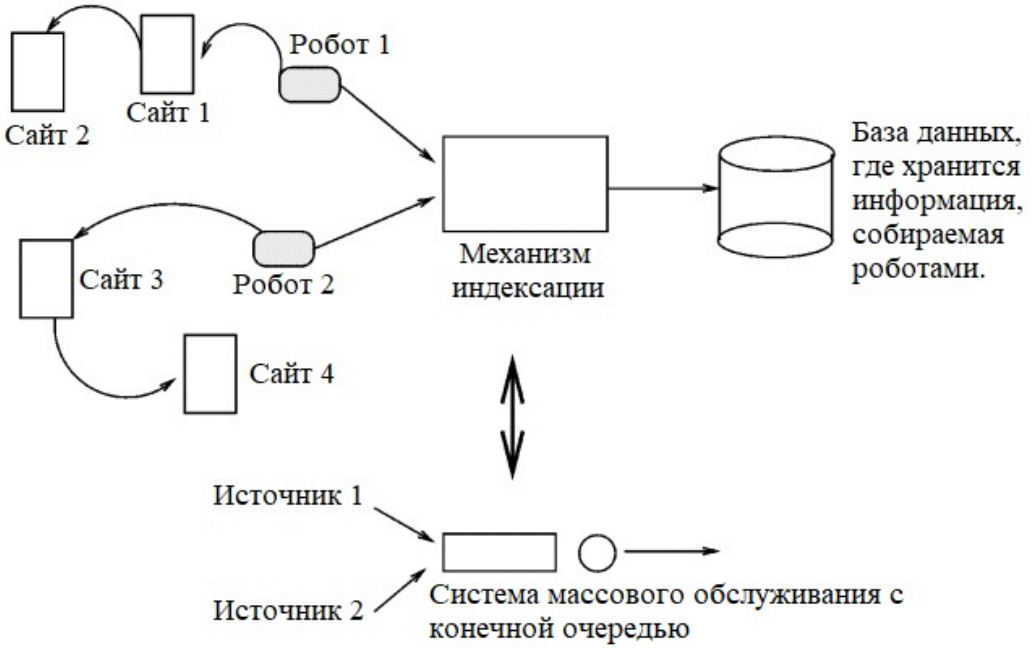

Рисунок 1 – Модель поисковой системы с двумя роботами.

Подраздел 1.2 посвящен моделированию информационно-поисковой системы с помощью одноприборной системы массового обслуживания с конечной очередью $M/M/1/K$ [7, 8]. Вместимость системы $K \geqslant 2$, включая требование, находящееся на обслуживании. В поисковой системе всего $N \geqslant 1$ роботов: каждый робот доставляет новые страницы согласно пуассоновскому процессу с интенсивностью $\lambda > 0$. Определено распределение вероятностей

длительности обслуживания $F(x) = P\{\sigma \leqslant x\}$, где σ - длительность обслуживания и $\bar{\sigma} > 0$ - математическое ожидание длительности обслуживания. Интенсивность обслуживания $\mu = 1/\bar{\sigma}$. В качестве характеристики качества функционирования системы определяется функция стоимости системы как взвешенная сумма двух слагаемых:

— вероятность простоя системы $P\{X = 0\}$, где X случайная величина, представляющая собой число требований в системе массового обслуживания с интенсивностью поступления требований λN и функцией распределения длительностей обслуживания F (в данном подразделе предполагается, что F - экспоненциальная функция);

— вероятность потери поступающих требований $P\{X = K\}$, то есть вероятность того, что число требований в системе массового обслуживания равно K.

Пусть $\rho := N\lambda/\mu > 0$, тогда функция стоимости определяется как

$$C(\rho, \gamma, K) := \gamma P\{X = 0\} + P\{X = K\}, \tag{1}$$

где γ - положительный весовой коэффициент вероятности простоя системы. Для нахождения вероятностей используется следующее утверждение:

Утверждение 1 [5]. Для любого $\rho > 0$

$$\mathbf{P}\{X = i\} = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}\rho^i & \text{для } i = 0, 1, ..., K, \\ 0 & \text{для } i > K. \end{cases} \tag{2}$$

Когда $\rho = 1$, $P\{X = i\} = 1/(K + 1)$ для $i = 0, 1, ..., K$.

Для того, чтобы найти оптимальное значение $N(\gamma, K)$ числа источников требований, необходимо уменьшить вероятность простоя системы $P\{X = 0\}$ и вероятность потери поступающих требований $P\{X = K\}$, то есть минимизировать функцию стоимости $C(\rho, \gamma, K)$. Пусть $\rho(\gamma, K)$ – оптимальная нагрузка системы. Решение задачи поиска оптимального значения $N(\gamma, K)$ приводится в следующем утверждении.

Утверждение 2 [5]. Для любых $\gamma > 0, K \geq 2$ пусть $N(\gamma, K)$ будет оптимальным числом источников требований. Тогда

$$N(\gamma, K) = arg \min_n C(n\lambda/\mu, \gamma, K), \tag{3}$$

где

$$n \in \{\lfloor \rho(\gamma, K)\mu/\lambda \rfloor, \lceil \rho(\gamma, K)\mu/\lambda \rceil\},$$

для любого натурального $x \lfloor x \rfloor$ (соответственно, $\lceil x \rceil$) означает округление вниз (соответственно, вверх) значения x.

Подраздел 1.3 посвящен моделированию информационно-поисковой системы с помощью системы массового обслуживания $M/G/1/K$ [7,8], у которой длительность обслуживания определяется функцией $\mathcal{F}(\theta) = \mathbf{E}[exp(-\theta\sigma)]$ – преобразование Лапласа-Стилтьеса распределения длительности обслуживания σ. Остальные параметры системы - такие же, как в предыдущем подразделе. Приводится выражение для стоимости системы $C(\rho, \gamma, K)$ [5].

По прямой аналогии с системой $M/M/1/K$ утверждение 2 дает приближение для оптимального числа источников требований $N(\gamma, K)$ в системе $M/G/1/K$.

Второй раздел «Управление числом поисковых роботов в информационно-поисковой системе» посвящен методу оптимального управления числом поисковых роботов в информационно-поисковой системе. В данном разделе ставится задача определения такой стратегии управления числом роботов, при которой минимизируется взвешенная сумма стационарных вероятностей простоя и потерь требований. Для решения задачи вводится марковский процесс принятия решений (далее МППР), с помощью которого описывается задача оптимального управления. Пространство состояний для МППР определяется следующим образом:

$$\boldsymbol{X} := \{(q, r, s), 0 \le q \le K, 0 \le r \le N, s = 0, 1, 2\}$$

$$-\{(0, 0, 2), (0, r, 0), (q, 0, 1), 0 \le q \le K, 0 \le r \le N\},$$

где q и r – длина очереди и количество активных роботов, соответственно, а s – тип (поступление требования, уход требования, фиктивное событие) момента принятия решения.

Суммарная интенсивность выхода системы из состояния определяется следующим образом:

$$\nu := \lambda N + \mu.$$

Множество допустимых действий $\boldsymbol{A}_x$, когда система находится в состо-

янии $x = (q, r, s) \in \boldsymbol{X}$:

$$\boldsymbol{A}_x = \begin{cases} \{0\}, & \text{если } s = 2, \\ \{1\}, & \text{если } (q, r, s) = (1, 0, 0), \\ \{0,\ 1\}, & \text{в остальных случаях,} \end{cases}$$

где $a = 1$, если принято решение активировать одного дополнительного робота, если таковой имеется, и $a = 0$, если принято решение оставить неизменным количество активных роботов.

При условии, что процесс находится в состоянии $x = (q, r, s)$ и что действие $a \in \boldsymbol{A}_x$ выполнено, одношаговая стоимость определяется как

$$c(x) = \gamma 1(q = 0) + \nu 1(q = K, s = 1) \tag{4}$$

и не зависит от a. Для $x \in \boldsymbol{X}$ одношаговые вероятности переходов $p_{x,x'}(a)$ определяются как

$$p_{x,x'}(a) = \begin{cases} \frac{\mu}{\nu} 1(q > 1), & \text{если } x' = (q - 1, min\{r + a, N\}, 0), \\ \frac{\lambda r}{\nu}, & \text{если } x' = (q - 1, min\{r + a, N\}, 1), \\ 1 - \frac{\mu 1(q>1) + \lambda r}{\nu}, & \text{если } x' = (q - 1, min\{r + a, N\}, 2), \end{cases} \tag{5}$$

если $s = 0, a = 0, 1$;

$$p_{x,x'}(a) = \begin{cases} \frac{\mu}{\nu}, & \text{если } x' = (min\{q + 1, K\}, r + a - 1, 0), \\ \frac{\lambda(r+a-1)}{\nu}, & \text{если } x' = (min\{q + 1, K\}, r + a - 1, 1), \\ 1 - \frac{\mu + \lambda(r+a-1)}{\nu}, & \text{если } x' = (min\{q + 1, K\}, r + a - 1, 2), \end{cases} \tag{6}$$

если $s = 1, a = 0, 1$;

$$p_{x,x'}(0) = \begin{cases} \frac{\mu}{\nu} 1(q > 0), & \text{если } x' = (q, r, 0), \\ \frac{\lambda r}{\nu}, & \text{если } x' = (q, r, 1), \\ 1 - \frac{\mu 1(q>0) + \lambda r}{\nu}, & \text{если } x' = (q, r, 2), \end{cases} \tag{7}$$

если $s = 2$. Все остальные вероятности перехода равны 0.

Оптимальная средняя стоимость θ и оптимальная стратегия π^* вычис-

ляются с использованием следующего рекуррентного алгоритма, известного как итерационный метод (метод итераций по значениям).

Утверждение 3 [6]. Пусть $\hat{x}$ - фиксированное состояние из $\boldsymbol{X}$ и $0 < \tau < 1$ - фиксированное число. Для $k \geq 0, x \in \boldsymbol{X}$ определим функции $(h_k; k \geq 0)$ как

$$h_{k+1}(x) = (1-\tau)h_k(x) + \tau(T(h_k)(x) - T(h_k)(\hat{x})),$$

где

$$T(h_k)(x) := c(x) + \min_{a \in \boldsymbol{A}_x} \sum_{x' \in \boldsymbol{X}} p_{x,x'}(a) h_k(x')$$

и $h_0(\hat{x}) = 0$, в остальных случаях h_0 является произвольным.
Тогда предел $h(x) = \lim_{k\to\infty} h_k(x)$ существует для каждого $x \in \boldsymbol{X}, \theta = \tau T(h)(\hat{x})$, и оптимальное действие $\pi^*(x)$ в состоянии x определяется выражением $\pi^*(x) \in argmin_{a \in \boldsymbol{A}_x} \sum_{x' \in \boldsymbol{X}} p_{x,x'}(a) h(x')$.

Третий раздел «Описание алгоритмов и программы для анализа функционирования информационно-поисковой системы» посвящен описанию алгоритмов, с помощью которых решаются задачи анализа, оптимизации, динамического управления числом поисковых роботов, и программной реализации разработанных алгоритмов на языке программирования *Python*.

В подразделе 3.1 описывается алгоритм решения задачи оптимизации для модели $M/M/1/K$. Для нахождения оптимальной нагрузки системы, минимизирующей стоимость функционирования системы, используется метод оптимизации функции одной переменной (метод золотого сечения). С использованием найденного значения определяется оптимальное число источников требований.

В подразделе 3.2 описывается алгоритм решения задачи оптимизации для модели $M/G/1/K$. Так же, как и в предыдущем подразделе, применяется метод оптимизации (метод деления отрезка пополам) для поиска оптимального значения нагрузки системы, которое, в свою очередь, используется при вычислении оптимального числа источников требований.

В подразделе 3.3 описывается алгоритм решения задачи динамического управления для модели $M/M/1/K$, реализующий итерационный метод поиска оптимальной стратегии управления.

Подраздел 3.4 содержит описание назначения, структуры и интерфей-

са программы. Программа для анализа функционирования информационно-поисковой системы с управлением числом поисковых роботов написана на языке *Python*. Программа включает следующие файлы:

— *StaticApproach* – файл, содержащий функции, реализующие методы анализа и оптимизации для систем массового обслуживания $M/M/1/K$ и $M/G/1/K$;

— *DynamicApproach* – файл, содержащий функции, реализующие метод динамического управления числом поисковых роботов;

— *main* – главный файл, реализующий ввод параметров системы и вычисление характеристик системы с помощью функций, реализованных в файлах *StaticApproach*, *DynamicApproach*.

Программа имеет оконный интерфейс. Присутствует возможность выбора типа модели системы, метода анализа для выбранной ранее модели, а также ввод необходимых параметров системы. В зависимости от выбранной модели, введенных параметров и выбранного метода анализа системы выбирается алгоритм для расчета характеристик системы и выводятся в окно программы результаты ее работы.

Четвертый раздел «Результаты исследования информационно-поисковой системы» посвящен описанию численных экспериментов, проведенных с помощью разработанной программы и связанных с исследованием функционирования информационно-поисковой системы с управлением числом роботов.

В экспериментах исследовалась:

— Зависимость оптимального числа источников требований $N(\gamma, K)$ от коэффициента вероятности простоя системы γ;

— Зависимость $N(\gamma, K)$ от вместимости системы K;

— Зависимость функции стоимости $C(\rho, \gamma, K)$ от нагрузки системы ρ;

— Зависимость оптимальной нагрузки системы $\rho(\gamma, K)$ от K;

— Зависимость функции стоимости $C(\rho, \gamma, K)$ при оптимальном числе источников требований и оптимальной средней стоимости θ при оптимальном динамическом управлении числом источников требований от вместимости системы K. Результаты данного эксперимента, представленные на рисунке 2, показали, что стоимость системы при динамическом управлении меньше, чем при оптимальном (статическом) количестве

источников требований.

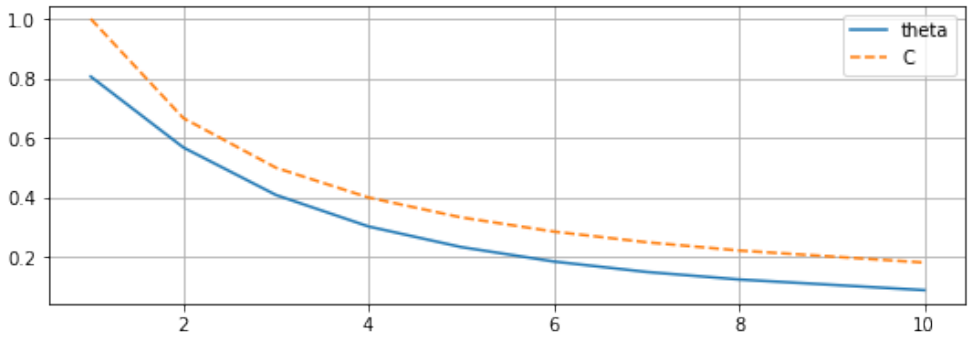

Рисунок 2 – Зависимость $C(\rho, \gamma, K)$ и θ от K.

ЗАКЛЮЧЕНИЕ

В выпускной квалификационной работе были изучены структура информационно - поисковой системы, использующей поисковых роботов, математические модели информационно-поисковой системы - системы массового обслуживания $M/M/1/K$ и $M/G/1/K$, методы анализа системы, оптимизации числа поисковых роботов и динамического управления числом роботов. Приведено описание разработанной программы для анализа модельных систем массового обслуживания.

Программа позволяет проводить анализ:

— системы $M/M/1/K$ без управления (с заданным числом источников требований);

— системы $M/G/1/K$ без управления (с заданным числом источников требований);

— системы $M/M/1/K$ с оптимальным числом источников требований;

— системы $M/G/1/K$ с оптимальным числом источников требований;

— системы $M/M/1/K$ с динамическим управлением числом источников требований.

С помощью программы были проведены исследования информационно-поисковой системы с управлением числом поисковых роботов. Изучались зависимости функции стоимости системы от различных параметров системы, в том числе при оптимальном статическом и динамическом управлениях числом роботов.

Основные источники информации:

1 Барашев, Д. В., Васильева, Н. С., Новиков, Б. А. Поисковая система [Электронный ресурс] // Большая российская энциклопедия [Электронный ресурс] : [сайт]. – URL : https://bigenc.ru/technology_and_technique /text/3151090 (дата обращения: 05.05.2022). – Загл. с экрана. – Яз. рус.

2 Chu, H., Rosenthal, M. Search engines for the World Wide Web: A comparative study and evaluation methodology // Proceedings of the Annual Meeting of the American Society for Information Science — 1996. — Vol. 33. — P. 127—135.

3 Tarakeswar, M. K., Kavitha, M. D. Search Engines: A Study // Journal of Computer Applications — 2011. – Vol. 4, № 1. — P. 29-33.

4 Coffman, E. G. Jr., Liu, Z., Weber, R. R. Optimal scheduling for web search

engines // Journal of Scheduling. – 1998. – P. 15-29.

5 Talim, J., Liu, Z., Nain, P., Coffman, E. G. Jr. Optimizing the Number of Robots for Web Search Engines // Telecommunication Systems – 2001. – Vol. 17, № 1,2. – P. 243-264.

6 Talim, J., Liu, Z., Nain, P., Coffman, E. G. Jr. Controlling the Robots of Web Search Engines // ACM Sigmetrics – 2001. – Vol. 29, № 1. – P.236-244.

7 Клейнрок, Л. Теория массового обслуживания / Л. Клейнрок; пер. И.И. Грушко; ред. В.И. Нейман. – М.: Машиностроение. – 1979. – 432 с.

8 Назаров, А. А. Теория массового обслуживания / А. А. Назаров, А. Ф. Терпугов. Томск: Издательство Научно-технической литературы, – 2004. – 229 с.