

ИМИТАЦИОННАЯ МОДЕЛЬ СИСТЕМЫ ИНФОРМАЦИОННОГО ПОИСКА В ОБЪЕКТНО-РЕЛЯЦИОННЫХ БАЗАХ ДАННЫХ

О.Н. Бородин, асп., асс. каф.

Вычислительной техники и прикладной математики

Тел.: 8(926) 374-07-38; E-mail: bor_oleg@rambler.ru

Московский государственный агроинженерный университет им. В.П. Горячкина

<http://www.msau.ru>

А.А. Евстифеев, к.т.н., совмест. каф. № 29

Тел.: 8 (910) 460-78-86; E-mail: evstipheev@mail.ru

Московский инженерно-физический институт (государственный университет)

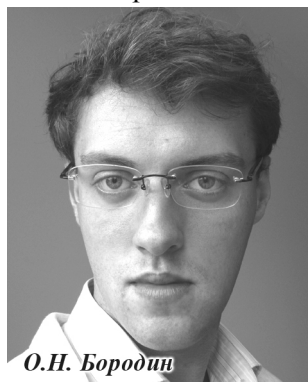
<http://www.mephi.ru>

Building and researching of system information retrieval based on simulation model in objective-relation database. Simulation model has been implemented using AnyLogic.

В государственных образовательных учреждениях увеличиваются массивы информационных документов, подлежащих хранению и систематизации. Информационными документами являются:

- справочные издания: словари, энциклопедии, справочники;
- научные труды: книги, монографии, статьи, материалы конференций, отчеты по НИР;
- нормативные документы: классификаторы, рубрикаторы, стандарты, законы, постановления, приказы;

В качестве инструмента хранения информационных документов используются объектно-реляционные базы данных. По мере накопления ин-



О.Н. Бородин

формационных документов в базе данных возникает проблема организации эффективного поиска информации по коллекциям документов за приемлемое время.

В качестве инструмента для хранения информационных документов используются объектно-реляционные базы данных. Такой выбор обусловлен следующими причинами:

- независимость информации, находящейся в базе данных, от файловой структуры операционной системы;
- поддержка объектных типов данных;
- возможность доступа к информации, содержащейся в базе данных, с помощью процедурных языков высокого уровня [1].

Решение проблемы поиска информации в объектно-реляционных базах данных базируется на достижениях лингвистики и информатики.

Под системой информационного поиска в объектно-реляционных базах данных понимается совокупность программных элементов, обеспечивающих решение задач информационного поиска.

Входными параметрами системы информационного поиска являются [2,3]:

1. Информационные документы. Информационным документом в базе данных является набор текстовой информации, сформированной на естественном языке. Информационный документ имеет определенное число информационных атрибутов, описывающих название документа, краткое содержание, полное текстовое содержание, представленное в двоичном файле определенного формата.

2. Информационные запросы. Информационным запросом является входное сообщение в систему информационного поиска, содержащее требование на выдачу информации. Задачей выполнения информационного запроса является выявление в множестве информационных документов, находящихся в базе данных, подмножества соответствующего информационной потребности пользователя.



А.А. Евстифеев

Входные параметры имеют следующие характеристики:

1. Количество параллельно обрабатывающихся в системе информационных документов и запросов.
2. Интенсивность поступления информационных документов и запросов в систему.
3. Размер информационного документа и запроса.

Значения входных параметров в потоке заявок на обслуживание носят случайный

характер, это приводит к тому, что в некоторые моменты времени в системе могут образовываться очереди на обработку. В результате возникает задача анализа системы информационного поиска как системы массового обслуживания в целях оценки временных характеристик обслуживания заявок на обработку информационных запросов и заявок на обработку информационных документов. Система информационного поиска как система массового обслуживания представлена на рис. 1.

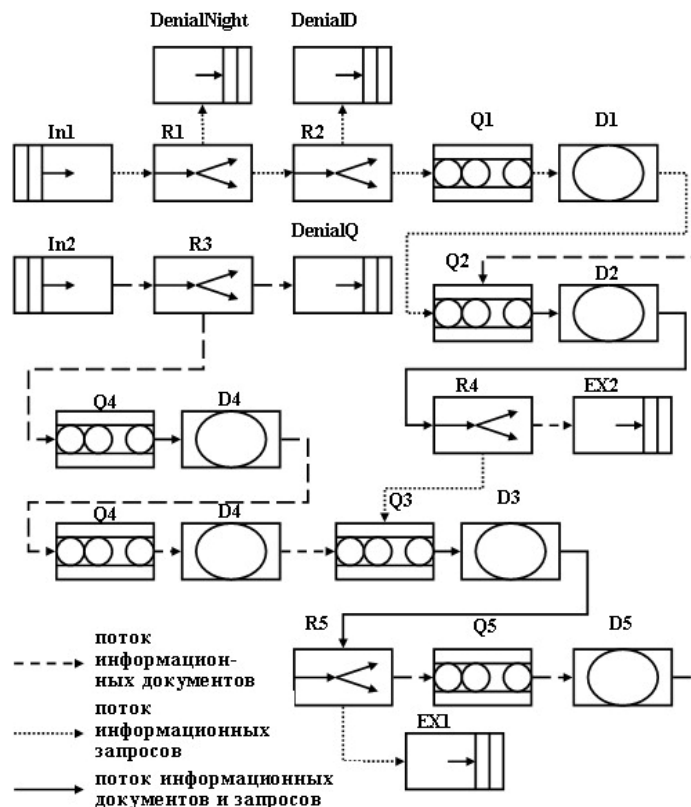


Рис. 1. Модель системы информационного поиска

Система массового обслуживания состоит из элементов Q и D типа (табл. 1). Элементы Q типа представляют очереди, в которых скапливаются информационные документы и запросы. Очередь сокращается в результате обработки заявки элементами D типа. Элементы D типа обрабатывают заявки с определенной производительностью [4].

Обслуживание заявок состоит из процессов обработки информационного документа, обработки информационного запроса.

Процесс обработки информационного документа состоит из следующих фаз:

1. Генерация информационного документа (элемент In1) с некоторой интенсивностью в рамках временного интервала равного 8-часовому рабочему дню.

2. Проверка степени загрузки каналов. В зависимости от состояния системы запрос поступает в очередь на обработку или отклоняется, если коэффициент загрузки каналов обслуживания $> 0,75$ (элементы – R2, denialD).

3. Формирование и обработка очереди, состоящей из информационных документов, направленных на обработку в программных интерфейсах информационной системы (элементы – Q1, D1).

4. Формирование и обработка очереди, состоящей из информационных документов, направленных на сохранение в табличном пространстве объектно-реляционной базы данных (элементы – Q2, D2).

5. Формирование и обработка очереди информационных документов, подлежащих индексации и лингвистическому анализу

(элементы – Q3, D3).

6. Регистрация обработанных информационных документов (элемент Ex1).

Процесс обработки информационного запроса состоит из следующих фаз:

1. Генерация информационного запроса (элемент In2).

2. Проверка степени загрузки каналов.

В зависимости от состояния системы запрос поступает в очередь на обработку или отклоняется, если коэффициент загрузки каналов обслуживания $> 0,75$ (элементы – R3, denialQ).

Таблица 1

Элементы модели системы информационного поиска

Элемент имитационной модели	Назначение
IN1	Входящий поток информационных документов в базу данных - случайного объема, действующий через фиксированные временные промежутки на протяжении полного времени выполнения эксперимента
R1	Фильтрация потока информационных документов, соответствующих 8-часовому интервалу из виртуального времени эксперимента
DenialNight	Регистрация информационных документов в соответствии с условиями элемента R1
R2	Фильтрация информационных документов при условии, если основные элементы системы не справляются с обработкой нового документа
denialD	Регистрация не обработанных в системе информационных документов
Q1,D1	Q1-очередь на обработку информационных документов в программных интерфейсах. D1-задерживает заявки на определенное время в соответствии со значением производительности элемента
Q3,D3	Q3-очередь на обработку информационных документов, требующих индексации и лингвистической обработки. D3 имитирует хранимую процедуру, выполняющую лингвистическую обработку текста, содержащегося в информационном документе и запросе
Q2,D2	Q2-очередь информационных документов, требующих сохранения в табличном пространстве базы данных. D2 имитирует операции сохранения информационных документов и операции выборки данных по коллекции документов в базе данных
R4	Разделение потока заявок на поток информационных документов и поток информационных запросов
EX1	Регистрация обработанных информационных документов
IN2	Входящий поток информационных запросов
R3	Выполняет фильтрацию информационных запросов при условии, если основные элементы системы не справляются с обработкой нового информационного запроса
DenialD	Регистрирует не обработанные в системе информационные документы
Q4,D4	Q4-очередь на обработку информационных запросов. D4 имитирует обработку информационных запросов программными интерфейсами в системе
Q5,D5	Q5-очередь поставленных на обработку базой данных информационных запросов. D5 имитирует построение информационных запросов с определенной производительностью.
R5	Разделение потока заявок на поток информационных документов и поток информационных запросов
EX2	Регистрация обработанных информационных запросов

3. Формирование и обработка очереди, состоящей из информационных запросов, направленных на обработку в программных интерфейсах информационной системы (элементы – Q4, D4).

4. Формирование и обработка очереди из информационных запросов, подлежащих лингвистическому анализу (элементы – Q3, D3).

5. Формирование и обработка очереди на построение информационного запроса (элементы – Q5, D5).

6. Выполнение информационного запроса по коллекции информационных доку-

ментов, находящихся в объектно-реляционной базе данных (элементы – Q2, D2).

7. Регистрация обработанных информационных запросов (элемент Ex2).

Все каналы обслуживания выполняют выборку из очереди по правилу FiFo, за исключением канала D3, выполняющего обработку заявок в режиме квантования времени. Потоки заявок не являются стационарными, а входной поток канала обслуживания D3 не является ординарным. В силу указанного принцип использования аналитических моделей, принятых в теории массового обслу-

живания, представляется проблематичным, так как принятое в них соглашение приводит к неадекватности аналитической модели.

В данной статье для решения задачи оценки характеристик системы информационного поиска рассматривается имитационная модель, реализующая принцип дискретно-событийного моделирования. В качестве инструментальной среды для реализации имитационной модели использована среда имитационного моделирования AnyLogic. Конструктивная библиотека Enterprise, встроенная в данную среду, имеет набор стационарных элементов, моделирующих

каналы обслуживания, очереди, источники заявок и маршрутизаторы заявок, что позволяет разрабатывать модели систем массового обслуживания любого типа [5]. Модель системы информационного поиска, построенная на основе стационарных элементов, соответствует системе массового обслуживания, приведенной на рис. 1.

Экспериментальные исследования системы информационного поиска на основе разработанной модели выполнены при предельном модельном времени 3000 ч и исходных данных, приведенных к табл. 2.

Таблица 2

Исходные данные имитационного эксперимента

Значение параметра	Параметр
3000	Имитируемое время работы системы, ч
3	Количество пользователей, осуществляющих ввод информационных документов в базу данных
0,5	Коэффициент количества загруженных информационных документов в час одним пользователем
10	Количество пользователей, выполняющих информационные запросы в базе данных
1	Коэффициент количества выполненных информационных запросов в час одним пользователем
1	Минимальное число страниц информационного документа
500	Максимальное число страниц информационного документа
200	Максимальное количество слов на странице информационном документе
1	Минимальная длина слова
7	Максимальная длина слова

Таблица 3

Результаты исследования системы информационного поиска

Показатели имитационного эксперимента	Характеристики альтернатив				
	1	2	3	4	5
Среднее время обработки информационного документа (t_1), с	10 8105	63 6,6	31 889	5 917	9 71
Среднеквадратичное отклонение DX по t_1	15 905,4	31 ,7	36 63,3	7 62,4	6 7,4
Среднее время обработки информационного запроса (t_2), с	6,8 3	0, 75	2, 55	0, 96	0, 77
Среднеквадратичное отклонение DX по t_2	1,0 5	0, 004	0, 24	0, 019	0, 006
Коэффициент обработанных документов, k_1	0,4 201	0, 999	0, 95	0, 98	0, 998
Коэффициент обработанных запросов, k_2	0,4 959	0, 998	0, 96	0, 99	0, 997

На основе разработанной модели выполнена оценка характеристик систем информационного поиска по пяти альтерна-

тивным вариантам реализации модуля лингвистической обработки текста (элемент D3). Рассмотрено пять альтернатив, по каж-

дой альтернативе получены показатели минимального ($t_{\min}$) и максимального ($t_{\max}$) времени обработки одной словоформы естественного языка.

1. Реализация элемента в виде динамически подключаемой библиотеки DLL, представленной компанией АОТ. Показатели альтернативы: $t_{\min}=0,2$ с, $t_{\max}=0,3$.

2. Использование встроенного в СУБД Oracle 10g решения. Показатели альтернативы: $t_{\min}=0,01$ с, $t_{\max}=0,01$.

3. ISPEL – свободно распространяемый морфологический словарь русского языка. Показатели альтернативы: $t_{\min}=0,1$ с, $t_{\max}=0,1$.

4. Russian Context Optimizer (RCO) – программный продукт, интегрируемый в СУБД Oracle 10g, позволяющий выполнять морфологический анализ текста с использованием технологии расширения информационного запроса дополнительным списком словоформ. Показатели альтернативы: $t_{\min}=0,04$ с, $t_{\max}=0,04$.

5. Модуль лингвистической обработки текста, разработанный на основе математической модели конечного детерминированного автомата. Показатели альтернативы: $t_{\min}=0,01$ с, $t_{\max}=0,02$.

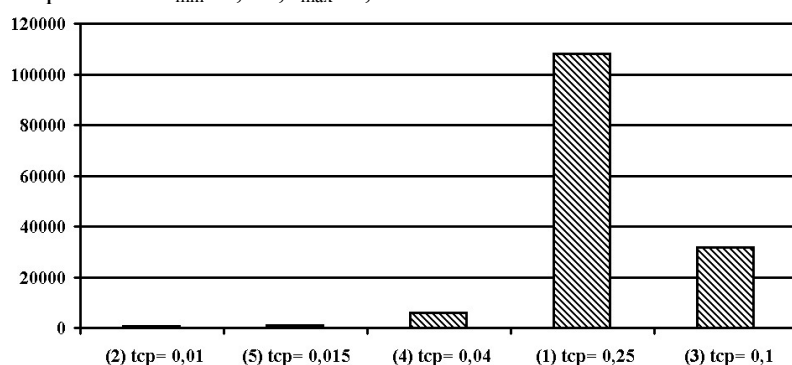


Рис. 2. Среднее время обработки информационного документа (t_1)

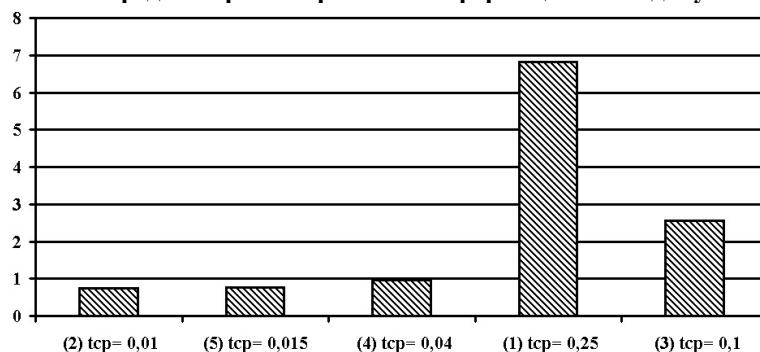


Рис. 3. Среднее время обработки информационного запроса (t_2)

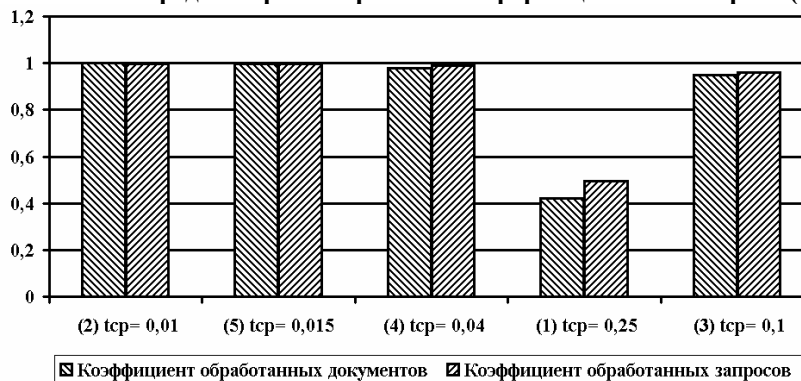


Рис. 4. Коэффициент обработанных документов (k_1) и запросов (k_2)

Результаты исследований системы информационного поиска по пяти альтернативным вариантам реализации элемента D3 приведены в табл. 3.

Зависимости характеристик системы информационного поиска от среднего вре-

мени обработки одной словоформы (t_{cp}) в модуле лингвистической обработки текста по пяти альтернативам приведены на рис. 2,3,4.

Результаты, полученные на основе имитационной модели, могут быть исполь-

зованы для анализа чувствительности систем информационного поиска и поддержки принятия решения по выбору оптимального (рационального) состава элементов системы информационного поиска. Возможность параметрической настройки разработанной

имитационной модели позволяет сократить затраты на проектирование систем информационного поиска в объектно-реляционных базах данных при принятии решения о выборе оптимального состава элементов.

Литература

1. Конноли Т., Бегг К., Страчан А. Базы данных: проектирование, реализация, сопровождение. Теория и практика. – 2-е изд. / Пер. с англ.: Уч. пос. – М.: Вильямс, 2000. – 1120 с.
2. Baezo-Yates R. Modern Information Retrieval / R.Baezo-Yates, B. Ribeiro-Neto. – ACM Press Addison Wesley, 1999. – 544 с.
3. Антонов А.В. Системный анализ: Учебник для вузов. – 2-е изд., стер. – М.: Высш. шк., 2006. – 454 с.
4. Шеннон Р. Имитационное моделирование систем: искусство и наука / Р. Шеннон. – М.: Мир, 1978. – 418 с.
5. Карпов, Ю.Г. Имитационное моделирование систем. Введение в моделирование с AnyLogic 5. – Ю.Г. Карпов. СПб: БХВ-Петербург, 2006. – 400 с.

СТРУКТУРА РЕАЛЬНОСТИ И ФЕНОМЕН ИНФОРМАЦИИ

К.К. Колин, д. т.н., проф.

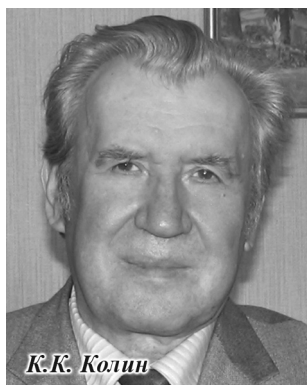
Тел.: 903-501-36-86; E-mail: kolinkk@mail.ru

Институт проблем информатики РАН

<http://www.ipiran.ru>

The philosophical essence of a phenomenon of the information as one of displays of fundamental properties of a reality of the world surrounding us is considered. Objective character of a phenomenon of the information and its communication with structure of a reality which possesses property of dualism is shown and simultaneously includes two basic components - a physical and ideal reality. Distinctive property of these components is their ability to mutual reflection, that, actually, and creates an opportunity of display of various aspects of a phenomenon of the information.

1. Осмысление сущности феномена информации – актуальная философская и научно-методологическая проблема XXI века



К.К. Колин

Анализ актуальных философских и научно-методологических проблем развития современной науки показывает, что одной из таких проблем является проблема осмысления природы и сущности информации.

По мнению многих авторитетных исследователей, информация представляет собой один из наиболее значимых и в то же

время загадочных феноменов окружающего нас мира. Попытки осмыслить сущность этого феномена предпринимались неоднократно многими учеными на протяжении нескольких десятилетий, начиная с середины XX века. Однако общепризнанных представлений о концептуальной природе информации в научной среде до сих пор еще не выработано, и поэтому эти попытки продолжаются и сегодня, на исходе первого десятилетия XXI века, который все более часто называют веком информации. Процессы становления информационного общества, а также те новые вызовы и общественно значимые потребности, которые являются характерными для развития цивилизации в XXI веке, также стимулируют новые попытки ученых к осмыслению концептуальной природы информации. Настоящая работа