Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.738.5-048.34

Холюков
Антон Андреевич

МЕТОДЫ ОПТИМИЗАЦИИ ЗАГРУЗКИ ИНТЕРНЕТ-РЕСУРСОВ

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 02 Системный анализ, управление и обработка информации

Научный руководитель

Муха Владимир Степанович
доктор технических наук,
профессор

Минск 2021

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Современные научные, образовательные центры имеют беспрецедентную возможность быстро и сравнительно дёшево знакомить региональное и мировое сообщество с результатами своей деятельности. Для этого достаточно создать один или несколько Web-серверов, на которых осуществлять электронную публикацию всех необходимых сведений. По такому пути пошли, например, в таком знаменитом научно-образовательном центре, как Стэнфордский университет (Stanford University), список серверов и сайтов которого содержит несколько сотен ссылок (http://www.stanford.edu/home/atoz). Но с ростом объемов опубликованной на всех Web-серверах информации посетителям становится трудно ориентироваться в получившемся сегменте информационного поля, а значит трудно находить путь к требуемым сведениям. Однако, при реализации поисковой системы (ПС), владельцы сегмента информационного поля сталкиваются с многочисленными проблемами выбора. Если осуществляется выбор готового продукта из большого количества коммерческих и свободно распространяемых программных комплексов ПС, то основной проблемой является отсутствие объективной (отличной от рекламных «фактов») информации по каждому рассматриваемому варианту (не наблюдается практика публикации числовых данных, которые характеризуют сегменты информационного поля и эффективность внедрения какой-либо системы). Поэтому потребности практики обуславливают необходимость разработки программных средств, которые позволят относительно легко (при минимальных требованиях) получить данные для объективного сравнения вариантов поисковых систем или некоторых стандартных модулей поисковых систем. Начиная самостоятельную разработку ПС оказывается трудно осуществить выбор наиболее эффективных внутренних алгоритмов (например, алгоритмов для подсистемы мониторинга состояния информационных ресурсов). Эта ситуация усугубляется тем, что в Internet присутствует большое количество публикаций (например, можно обратиться к материалам международных ежегодных специализированных конференций: http://www2003.org/, http://www.iadis.org/icwi2003), в которых предлагаются методы совершенствования различных аспектов функционирования ПС. Помочь сделать выбор может использование методов имитационного моделирования для предварительного исследования эффективности каждого из альтернативных алгоритмов. Однако в проанализированных работах наблюдается явная нехватка не только готовых моделей, но и подходов к их построению и программной реализации.

Цель работы состоит в разработке моделей и программного комплекса на их основе, позволяющих получить данные для проведения исследований относительной эффективности функционирования различных вариантов системы мониторинга. Для достижения поставленной цели решаются следующие задачи:

– выявить набор критериев эффективности, которые могут охарактеризовать любой вариант системы мониторинга;

– разработать математическую модель системы мониторинга;

– разработать имитационные модели системы мониторинга;

– разработать программный комплекс, который предназначен для проведения дискретно-событийного имитационного моделирования систем мониторинга;

– получить экспериментальное подтверждение практической применимости разработанного программного комплекса.

Методы исследования. При решении поставленных задач в работе использованы элементы математического и имитационного моделирования, а также методы и средства объектно-ориентированного программирования и баз данных.

Научная новизна. К новым результатам диссертации можно отнести:

– предложенную математическую и имитационные модели процесса мониторинга информационного поля Internet;

– предложенную сенсорную технологию осуществления мониторинга;

– разработанный программный комплекс имитационного моделирования системы мониторинга сегмента информационного поля;

Практическая ценность работы заключается в возможности использования полученных научно-технических результатов при проектировании, эксплуатации систем мониторинга.

Основные положения, выносимые на защиту:

– модели систем оптимизации загрузки интернет-ресурсов в поисковых системах, которые принадлежат различным классам в классификации стратегий мониторинга;

– SimCOSAR - программный комплекс дискретно-событийного имитационного моделирования систем мониторинга;

– Результаты проведённых имитационных компьютерных экспериментов.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из введения, четырех глав, заключения, библиографического списка и четырёх приложений.

Во введении обоснована актуальность темы диссертации, изложена цель и задачи исследования, научная новизна и практическая значимость, представлены основные положения, выносимые на защиту.

В первой главе фиксируются значения используемых в работе терминов, рассматриваются основные проблемы, которые возникают перед создателями и владельцами систем мониторинга, предлагается классификация алгоритмов и стратегий мониторинга, рассматриваются критерии эффективности систем мониторинга.

В работе под словами «информационный ресурс» (ИР) понимается файл (последовательность байт), который расположен на сервере. Внутренняя структура и тип информации (HTML, RTF, JPEG, SWF, AVI и т.п.) файла значения не имеют.

Объединение в рамках одного множества всех доступных информационных ресурсов (расположенных на некотором множестве Web-серверов) составляет информационное поле Internet. Тогда сегментом информационного поля Internet будет некоторая часть всего информационного поля Internet.

Проблемы, которые непосредственно влияют на принимаемые решения при проектировании систем мониторинга информационного поля можно разделить на два класса. В первый класс относятся проблемы, связанные с «природой» ИР:

– большой суммарный объем (байт) всех ИР;

– распределение НР между узлами сети;

– разное время существования ИР. Документы или файлы могут быть легко добавлены и также легко удалены в Web;

– динамичность изменения содержимого ИР;

– «доступность» ИР. Имеются в виду различные качество и стабильность функционирования коммуникационных связей между распределёнными узлами и сегментами Internet.

– разнородность ИР. Имеются в виду, например, различные форматы файлов или применяемые естественные языки.

– различное «качество» ИР. Имеются в виду проблемы профессионализма создателей, а также правовые и морально-этические проблемы.

– «скрытость» ИР. Здесь имеются в виду ИР для доступа, к которым нужно пройти сложную процедуру регистрации или сформировать некий запрос с помощью специальной формы.

– различная «популярность» и уровень «полезности» ИР

Во второй класс относятся проблемы нагрузки на задействованные аппаратные ресурсы:

– минимизация нагрузки на информационный источник

– минимизация нагрузки на каналы связи

– оптимизация нагрузки на модули сбора и накопления информации.

Все предлагаемые и используемые алгоритмы и стратегии можно распределить по двум конкурирующим концепциям: «роботов», «сенсоров».

Концепция «роботов». Основным признаком такой системы мониторинга является сервер или кластер серверов, на которых выполняется специальный программный код (именуемый в литературе «сетевым роботом», пауком, краулером, индексатором) - «робот». Причём верно то, что:

– робот постоянно занят скачиванием ИР из доступного сегмента информационного поля;

– очередь ИР на скачивание организуется согласно некоторой внутренней стратегии, которую задают разработчики;

– размер доступного роботу сегмента информационного поля увеличивается за счёт ручного добавления адресов ИР или за счёт обнаружения новых адресов в уже известных ИР.

Имеется несколько наиболее популярных направлений решения проблем нагрузок:

– Разработка специализированных роботов (имеется в виду жёсткие правила отбора ИР для их добавления в обрабатываемый сегмент информационного поля, например, когда обрабатываются ИР только по музыкальной тематике).

– Вычисление «рейтинга» ресурсов (имеется в виду некоторая композиция значения частоты изменений ИР и его вероятной «полезности»).

– Более полное использование стандартных средств HTTP-протокола (запрос HEAD, поле If-Modified-Since) - далее в работе соответствующие роботы именуются «модифицированные роботы».

Концепция «сенсоров». Системы мониторинга данного типа отличает потребность в установке на все принадлежащие информационному полю Web-сервера специального программного модуля, на который ложится полная ответственность за обнаружение изменений в уже известных ИР, а также частичная ответственность за обнаружение новых ИР. Причём верно то, что

сведения о новом состоянии информационного источника передаются на головной узел (или всем заинтересованным в этой информации узлам Internet) системы мониторинга, где на основе этих сведений предпринимаются действия (например, принимается решение о скачивании того или иного ИР) по актуализации представления об информационном поле.

Критерии эффективности. В качестве основных были выбраны два наиболее общих и объективных числовых критерия эффективности для любых вариантов системы мониторинга. Во-первых, «свежесть» (freshness) накопленной информации, один из возможных вариантов вычисления которой имеет вид:

$$F_p(S;t) = \frac{100}{N}\left(N - \sum_{i=1}^{N} F_N(e_i;t)\right)$$

Где $F_p(S;t)$ – свежесть, измеряемая в процентах; N – общее количество ИР; e_i – ИР, которые уже хранятся в базе данных системы мониторинга; $\boldsymbol{S} = \{\boldsymbol{e_1}, \ldots, \boldsymbol{e_N}\}$ – база данных системы мониторинга; $F_N(e_i;t)$ – свежесть элемента e_i в момент времени t, которая вычисляется следующим образом:

$$F_{\mathcal{N}}(e_i;t) = \begin{cases} 0, \textit{ если } e_i \textit{ не требует обновления в момент } t \\ 1, \textit{ если } e_i \textit{ устарел} \end{cases}$$

Во-вторых, суммарный объём ***V(t)*** (в байтах) всех ИР, которые система мониторинга перекачала со стороны информационных источников на сторону головного узла мониторинга с начала работы до момента наблюдения t.

Было отмечено, что объективное сравнение значений выбранных критериев эффективности возможно только в случае равенства условий, в которых получены сравниваемые значения.

Вторая глава полностью посвящена разработке математической модели процесса мониторинга и имитационных моделей системы мониторинга, которые могут служить в качестве примера удобного, достаточно строгого и единообразного подхода к описанию систем мониторинга, а также станут основой программных средств, реализуемых в данной работе.

Математическая модель. Информационное поле состоит из N различных информационных ресурсов. Процесс мониторинга – это меняющийся в дискретном времени набор состояний. Состояние в момент времени t - это тройка <R,R', π>, где

R = (r1,r2,..., rN), ri - матрица числовых характеристик информационного ресурса (к примеру, признак свежести, объём, идентификатор текущего состояния и пр.) с номером i;

R' = (r'1,r'2,..., r'N), r'i - матрица числовых значений критериев эффективности Freshness и Sumsize информационного ресурса с номером i;

π=(π1,..., πN), πi - текст программы записанной на языке высокого уровня, преобразующая по некоторому алгоритму матрицу ri в матрицу r'i

Алгоритмы преобразования сначала описаны в имитационных моделях, а затем реализованы в виде программ на языке Python.

Имитационные модели. Были выбраны два варианта («обычный робот» и «на сенсорах») системы мониторинга, для каждого из которых в диссертационной работе предложена своя имитационная модель. Модели представляют собой отображение причастных к мониторингу объектов и алгоритмов их функционирования.

Одинаковым для обеих моделей является, во-первых, то, что основные объекты в них – объекты типа «информационный ресурс», которые характеризуются значениями следующих свойств:

– размер содержимого (количество байт);

– идентификатор текущего состояния (рассматривается конечный набор допустимых значений, каждое из которых относится к группе «ИР доступен» или «ИР недоступен»);

– признак того, что серия последних изменений стала известна системе мониторинга (истина - ложь).

Второй одинаковой чертой моделей является наличие объектов типа источник изменений (ИИ) и источник запросов (ИЗ), каждый из которых описывается потоком изменений (ПИ) или потоком запросов (ПЗ) соответственно. В свою очередь, поток изменений характеризуется законом распределения $\mathbf{H_c(x)}$ времени ($T_{изм}$) появления следующего изменения (смены состояния ИР или смены содержимого ИР) и законом распределения $\mathbf{G_v(\omega)}$ относительной частоты v появления определённого (одного из возможных вариантов) «изменения» $И_{нов}$, а поток запросов характеризуется законом распределения $\mathbf{H_q(x)}$ времени ($T_{запр}$) появления следующего запроса. Каждому ИР сопоставляется собственная пара ИИ и ИЗ.

Различие моделей проявляется в объектах, которые воспроизводят элементы системы мониторинга:

– В модели «робота» систему мониторинга воспроизводят два объекта: «робот» и «репозиторий». Причём в модели описан робот, который реализует стратегию последовательного скачивания всех известных ИР.

– В модели «сенсоров» систему мониторинга воспроизводят объекты типа «сенсор», «робот» и «репозиторий». Причём в модели реализуется робот, который немедленно начинает скачивание ИР сразу после поступления от соответствующего «сенсора» уведомления о найденном изменении состояния ИР или изменении содержимого ИР. Для простоты рассмотрения считалось, что каждому экземпляру объекта ИР соответствует один экземпляр объекта «сенсор».

«Репозиторий» в обоих случаях выполняет одну и ту же роль - хранит скачанную роботом информацию об информационных ресурсах.

Для моделирования процесса скачивания в обеих моделях используется закон распределения Hg(x) времени скачивания ИР, с помощью которого определяется момент завершения процесса скачивания этого ИР - Тскач. Для имитации процесса посылки уведомления сенсором определяется время Тувед когда сигнал уведомления «дойдёт» до робота (время Тувед имеет свой закон распределения Ha(x) чтобы в последствии можно было учитывать такие случаи как, например, отложенное уведомление или загруженность Web-сервера).

Третья глава полностью посвящена разработке и описанию программного комплекса дискретно-событийного имитационного моделирования работы вариантов системы мониторинга, которые используют роботов, модифицированных роботов, сенсоры.

В качестве программной платформы для реализации компьютерных моделей были выбраны следующие программные продукты. Язык программирования Python 2.2, средства баз данных MySQL 4.0.13, python-модуль SimPy 1.3.

Python-модуль SimPy представляет собой библиотеку классов для программирования моделей, которые основаны на дискретных событиях (классы реализуют ведение календаря событий).

Программный комплекс SimCOSAR. Были выделены следующие основные операции, которые составляют процесс моделирования: создание набора ресурсов, создание журналов изменений и запросов ресурсов, функционирование системы мониторинга. В результате были разработаны следующие модули:

– SimPages.py - задача модуля заключается в генерации (равномерный закон распределения задаёт начальный размер каждого ИР) набора ИР.

– SimChanges.py - задача модуля заключается в генерации (интервал времени между последовательными событиями определяет экспоненциальный закон распределения) истории изменений ИР определённого набора. Тип нового изменения определяется с помощью заданных относительных частот появления шести предопределённых типов изменений.

– SimQueries.py - задача модуля заключается в генерации (интервал времени между последовательными событиями определяет экспоненциальный закон распределения) истории запросов ИР определённого набора.

– SimRobRoute.py - модуль занимается тем, что «привязывает» каждый конкретный ИР к одному из роботов. Используемый алгоритм «привязки» априори подразумевает наличие упорядоченной нумерации у ИР, которые созданы модулем SimPages.py.

– SimRobot.py - модуль позволяет осуществлять имитацию функционирования вариантов системы мониторинга, которые построены с использованием «обычных роботов».

– SimRobotM.py - имитация работы варианта системы мониторинга, которая использует «модифицированных роботов». За основу модуля был взят код модуля SimRobot.py и соответствующим образом модифицирован.

– SimSensor.py - модуль позволяет осуществлять имитацию функционирования варианта системы мониторинга, которая использует «сенсоры».

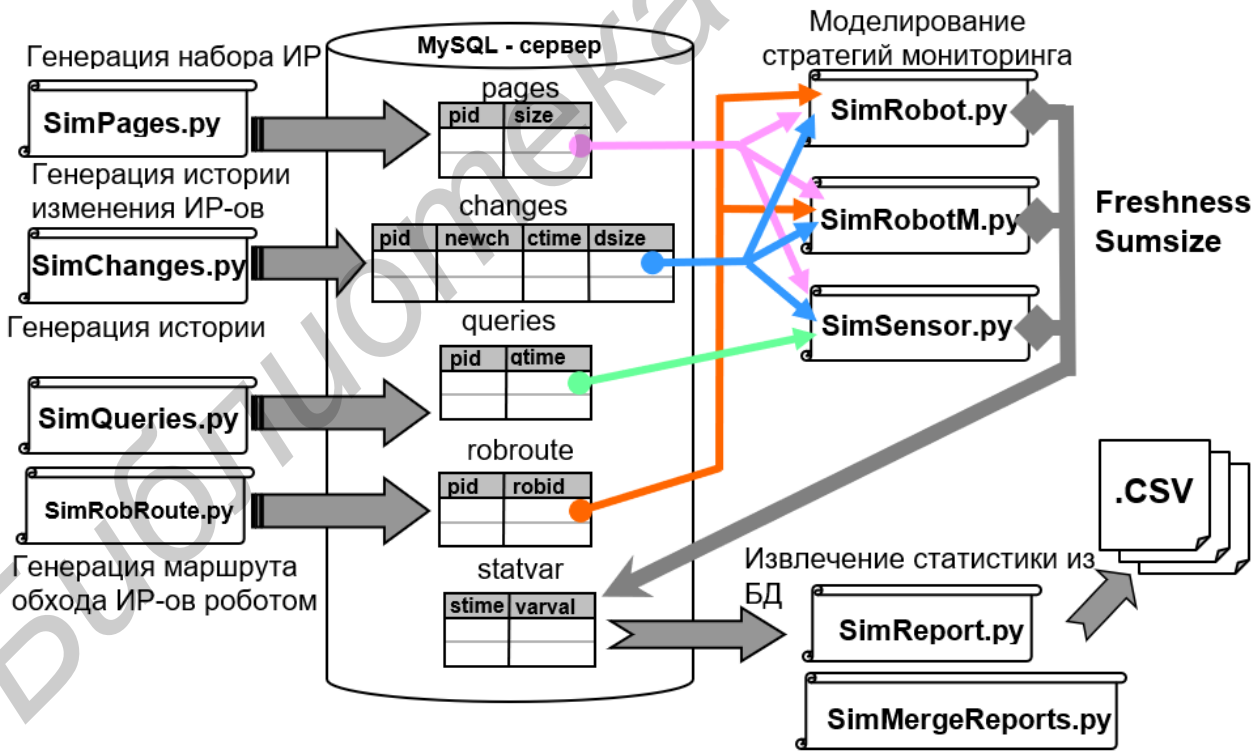

Рисунок 1 – Структура комплекса SimCOSAR

Также имеются два модуля, которые предназначены для извлечения из баз данных значений статистических переменных и их последующей записи в файлы формата CSV:

– SimReport.py - извлекает из баз данных значения, которые накоплены в ходе одного опыта.

– SimMergeR.eports.py - извлекает статистику нескольких опытов (вместо нескольких CSV-файлов получаем один файл - первый столбец содержит время замера, а в остальные столбцы попадают результаты замеров одной переменной в нескольких опытах).

Четвёртая глава посвящена подробному описанию эксперимента, который был реализован с использованием программного комплекса SimCOSAR. Целью эксперимента являлось получение новых знаний об относительной эффективности нескольких вариантов системы мониторинга в широком диапазоне рабочих нагрузок (т.е. получение подтверждения адекватности разработанных моделей и программ).

Планирование эксперимента. Проведён анализ имеющихся в моделях переменных, в результате которого выделены основные, второстепенные, третьестепенные переменные отклика и основные, второстепенные, третьестепенные влияющие факторы. К основным переменным отклика отнесены переменные:

– Количество «устаревших» ресурсов (Needfresh, шт.).

– Свежесть (Freshness, %).

– Объём скачанной информации (Sumsize, байт).

Для основных влияющих факторов выбраны уровни изменения:

– Количество наблюдаемых страниц (шт.) - 100000,200000,300000.

– Продолжительность эксперимента (ед., 1 единица = 100 мсек) - 8640000 (10 дней), 17280000 (20 дней).

– Интенсивность изменений (шт./длительность периода) -1,5 , 10 (для 10 дней); 2,10,20 (для 20 дней).

– Интенсивность запросов (шт./длительность периода) - 1, 50, 100 (для 10 дней); 2,100,200 (для 20 дней).

– Вариант системы мониторинга - выбрано пять вариантов, которые определяются использованием в процессе мониторинга:

- одного «обычного» робота (циклически скачивает подряд все ресурсы) - вариант № 1;
- одного «модифицированного» робота (циклически посылает всем ресурсам специальный HTTP запрос и при необходимости скачивает изменившийся ресурс) - вариант №2;
- трёх обычных роботов - вариант №3;
- трёх модифицированных роботов - вариант №4;
- «сенсоров» - вариант №5.

В качестве дополнительных условий в эксперименте считалось, что, во-первых, каждый вариант системы мониторинга работает на единственном компьютере, а во-вторых, вычислительная мощность используемых машин одинакова у всех вариантов.

В результате применения описанного на стадии стратегического планирования алгоритма получено 270 уникальных сочетаний основных влияющих факторов, каждое из которых описывает ситуацию, когда мониторингом 1 из 54 уникальных «информационных полей» занимается 1 из 5 уникальных вариантов системы мониторинга.

Основным результатом тактического планирования стало предложение проводить этап создания 54 «информационных полей» только один раз вместо положенных пяти (так как имеется пять вариантов системы мониторинга).

Реализация. Приведена информация о стадии подготовки программного комплекса SiraCOSAR к проведению эксперимента. Эти сведения могут служить в качестве наглядной иллюстрации того, как от плана эксперимента перейти к настроенному на него программному комплексу.

В работе представлены подробные сведения, которые характеризуют использованные вычислительные мощности и затраченное время на реализацию эксперимента. Эта информация позволяет оценивать ресурсы, которые могут потребоваться для проведения экспериментов по другим планам. Например, указаны следующие факты:

– Один прогон эксперимента по составленному плану моделирует 489888000 секунд или 5670 виртуальных дней (каждый из пяти сравниваемых вариантов работал на 54 «информационных полях», суммарное виртуальное время которых равняется 69984000 секундам или 810 виртуальным дням).

– Благодаря распределению вычислений по нескольким машинам реализация одного прогона заняла примерно два месяца.

– Если бы эксперимент осуществлялся без остановок и в каждый момент его реализации, выполнялась бы только одна операция на одной единственной машине, то продолжительность одного прогона эксперимента равнялась бы 30817758 секундам (реального времени) или примерно 357 дням.

Анализ результатов эксперимента. Было показано, что:

– получаемые значения основных (Freshness, Sumsize) критериев эффективности системы мониторинга действительно зависят (чувствительны к изменению) от значений выбранных (ChangeIntensity, RequestIntensity, PageCount, ModelTime) факторов. Для этого использовались финальные значения критериев эффективности, таблица коэффициентов корреляции);

– получаемые значения переменных Freshness и Sumsize действительно помогают оценить относительную эффективность нескольких (даже

принадлежащих разным концепциям) вариантов системы мониторинга. Для этого был построен частный рейтинг эффективности вариантов, которые принадлежат концепции роботов. Также было показано, в каких случаях сенсорный вариант системы мониторинга занимает лидирующее положение в общем рейтинге вариантов, а в каких явно уступает.

В заключении сформулированы основные результаты диссертационной работы и намечены дальнейшие пути развития исследования.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

При выполнении диссертационной работы получены следующие основные результаты.

1. Предложена новая классификация стратегий мониторинга информационных полей Internet. В качестве классифицирующего признака выбрана информация о месторасположении модуля системы мониторинга, который отвечает за обнаружение изменений в информационных ресурсах.

2. Использование введённой классификации позволило предложить новый (не упоминавшийся в публикациях ранее) алгоритм мониторинга, который основан на «сенсорах» (программных модулях), которые «слушают» трафик Web-сервера.

3. Выявлены (в ходе анализа публикаций) критерии эффективности, которые пригодны для сравнения любых вариантов системы мониторинга: «свежесть» накопленной системой мониторинга информации и объём данных, которые передаются с Web-серверов на головной сервер мониторинга.

4. Предложена математическая модель процесса мониторинга.

5. Разработаны имитационные модели двух вариантов системы мониторинга. Во-первых, системы, которая использует одного обычного робота для циклического скачивания всех информационных ресурсов информационного поля. Во-вторых, системы, которая использует сенсоры для обнаружения изменений состояния информационных ресурсов и которая после поступления от сенсора «сигнала тревоги» немедленно скачивает изменившийся ИР.

6. На базе языка программирования Python, БД MySQL, библиотеки классов дискретно-событийного моделирования SimPy и имитационных моделей разработан программный комплекс SimCOSAR для компьютерного моделирования работы вариантов системы мониторинга, которые используют роботов, «модифицированных роботов», сенсоры.

7. В ходе компьютерного экспериментирования с комплексом SimCOSAR получены числовые данные, после анализа которых:

- признано, что они адекватно отражают характеристики протекающих процессов (так как очевидно, что смоделированные ситуации можно воссоздать в практических условиях);
- получены новые знания (например, составлен рейтинг вариантов) об относительной эффективности пяти вариантов системы мониторинга в широком (54 ситуации) диапазоне рабочих нагрузок - сенсорную систему мониторинга рекомендуется внедрять для наблюдения за ИР, которые имеют

интенсивность посещений большую, чем интенсивность изменений, а для наблюдения за ресурсами, у которых интенсивность изменений превышает интенсивность посещений, рекомендуется использование систем мониторинга основанных на модифицированных роботах;

- признано, что комплекс применим (следовательно, применимы и модели, на которых он основан) для получения числовых данных, которые необходимы при сравнении различных вариантов системы мониторинга.

8. Полученные в эксперименте данные помогли принять решение о начале реализации сенсорной системы для мониторинга информационного поля Internet университета.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1-А Холюков, А.А. Потоковая передача данных в сети Интернет/ А.А. Холюков// Студенческий. СибАК. – 2021. – №2(130) – С.70-76.