

Кыргызский Национальный Университет им. Ж. Баласагына

Институт Интеграции Международных Образовательных Программ



КЫРГЫЗСКО-АМЕРИКАНСКИЙ ФАКУЛЬТЕТ
КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ И ИНТЕРНЕТ

УДК 681.3 (575.3) (043)

Похилько Андрей Федорович

**Анализ оптимальности информационных структур в
системах гипертекстовых документов**

**Диссертация на соискание квалификационной академической степени магистра
техники и технологии**

Направление: 552800 Информатика и вычислительная техника

Специализация: Компьютерные информационные системы и Интернет

Диссертация рекомендована к защите

Декан КАФ-Интернет:
член-корр. НАН КР,
д.т.н., проф.

_____ Бримкулов У.Н.

« ____ » июнь 2006 г.

Научный руководитель:
Академик НАН КР,
д.т.н., проф.

_____ Живоглядов В.П.

Нормоконтролер:
ст. преподаватель

_____ Большакова Т.Н.

Кыргызский Национальный Университет им. Ж. Баласагына
Институт Интеграции Международных Образовательных Программ (ИИ МОП)
Кыргызско-Американский Факультет
Компьютерных технологий и ИНТЕРНЕТ (КАФ-ИНТЕРНЕТ)

УТВЕРЖДАЮ
Декан КАФ-Интернет,
член-корр. НАН КР, д.т.н., проф.

_____ Бримкулов У.Н.
«___» июня 200__ г.

ЗАДАНИЕ НА МАГИСТЕРСКУЮ ДИССЕРТАЦИЮ

СТУДЕНТУ: **ПОХИЛЬКО Андрею Федоровичу**

ТЕМА МАГИСТЕРСКОЙ ДИССЕРТАЦИИ: **"Анализ оптимальности информационных структур в системах гипертекстовых документов"**

Утверждена приказом по ИИМОП от «_____» «_____» «200__» г. № _____

Срок сдачи студентом законченной работы «_____» «_____» 200__ г.

ИСХОДНЫЕ ДАННЫЕ К МАГИСТЕРСКОЙ ДИССЕРТАЦИИ:

1. *"Алгоритмы решения некоторых теоретико-графовых задач", А.Чернобаев*
2. *Статья "Показатель суммарной содержательной ценности информационных блоков в задаче оптимизации ссылочных структур гипертекстовых документов", Полковников Е.В.*
3. *Фактические данные о системе гипертекстовых документов веб-сервера <http://www.online.kg/>*

СОДЕРЖАНИЕ ДИССЕРТАЦИИ (перечень вопросов подлежащих разработке):

1. *Определение области применения*
2. *Разработка метода оценки структуры системы документов*
3. *Разработка инструментального средства для сбора информации о системах гипертекстовых документов*
4. *Сбор фактической информации на сервере [online.kg](http://www.online.kg/)*
5. *Анализ типовых информационных структур*
6. *Анализ фактических данных*
7. *Выводы и наблюдения*

ДАТА ВЫДАЧИ ЗАДАНИЯ «_____» «_____» 200__ г.

РУКОВОДИТЕЛЬ: **Живоглядов В.П.**

ЗАДАНИЕ ПРИНЯЛ К ИСПОЛНЕНИЮ: «_____» «_____» 200__ г.

РЕФЕРАТ

Магистерская диссертация: 44 с., 8 рис., 18 источников, 1 приложение.

Ключевые слова: гипертекст, структура, теория графов, критерии, оптимизация

Объектом исследования являются критерии оптимальности ссылочной структуры гипертекста.

Цель работы – разработка метода анализа оптимальности ссылочных структур в гипертексте и их оптимизации на основе формальных критериев.

Современные гипертекстовые системы характеризуются большими количествами документов и ссылок между ними. В процессе разработки гипертекст имеет четкую иерархическую структуру, но при реализации появляется множество перекрестных ссылок, приводящее к деструктуризации гипертекста и потере контроля над качеством системы. При использовании такой системы пользователь может быть дезориентирован и прекратить чтение гипертекста, что является крайне нежелательным для разработчика.

В первой части работы проведено исследование опыта анализа гипертекстовых систем и роли математической теории графов как средства формализации гипертекста. Во второй части сформулирована проблема потери контроля над структурой разрастающихся гипертекстовых систем, выделен ряд критериев формальной оценки качества ссылок в гипертексте.

В третьей части приведен пример программы, реализующей разработанный метод на языке T-SQL, и опробовано ее применение для анализа реальных гипертекстовых систем. В результате подчеркнута необходимость итерационного применения метода и целесообразность определенной последовательности использования критериев при оптимизации гипертекста.

РЕФЕРАТ

Магистердик диссертация: 44 барак, 8 сүрөт, 18 булак

Негизги колдонулуучу сөздөр: гипертекст, структура, граф теориясы, критерийлер, оптимизация

Изилдөөнүн объекти: Гипертексттин таянуу структурасынын оптималдуулугун өлчөөдөгү критерийлер.

Жумуштун максаты: гипертексттеги таянуу структураларынын оптималдуулугун анализдөө ыкмасын иштеп чыгаруу жана аларды формалдуу критерийлердин негизинде оптимизациялоо.

Заманбап гипертексттик системалар иш кагаздарынын чоң көлөмү менен жана алардын ортосундагы көп сандагы таянуулар менен аныкталат. Иштетүү процессинин ичинде гипертекст иерархиялык структурага ээ, бирок ишке ашыруу мезгилинде көптөгөн кесилишкен таянуулар пайда болот да, гипертекстти деструктуризацияга алып барууга жана сапатын башкаруу мүмкүнчүлүгүнө тоскоол кылат. Мындай системаны колдонгон иштетүүчү адашып кетет да, гипертекстти окуусун токтотот.

Жумуштун биринчи бөлүгүндө гипертексттик системанын анализи аныкталган, жана ошондой эле графалардын гипертекстти формализациядагы математикалык ролу көрсөтүлгөн. Ал эми экинчи бөлүгүндө болсо, гипертексттик системанын таралуу тармагын башкаруу мүмкүнчүлүгүн жоготуу маселеси көрсөтүлгөн. Ошондой эле гипертексттеги таянуулардын сапатын формалдуу түрдө сыноо ыкмасы баяндалган.

Үчүнчү бөлүктө мисал катары T-SQL тилинде иштетилген ыкманы колдонуучу программа тандалып алынып, аны реалдуу гипертексттик системаларын анализдөө максатында колдонуусу көрсөтүлгөн. Изилдөөнүн негизинде гипертекстти оптимизациялоодо ыкманын итерациялдуу колдонулушу жана критерийлерди белгилүү удаалаштыкта колдонулушунун зарылдыгы белгиленген.

ABSTRACT

Master's thesis: 44 p., 8 pic., 18 sources, 1 enclosure

Keywords: hypertext, structure, graph theory, criteria, optimization

The object of the research is the criteria of the hyperlink structure optimization of the hypertext.

Purpose of the work – the development of the analysis method of the hyperlink structures optimization in the hypertext and their optimization based on the formal criteria.

Present-day hypertext systems could be described by large amount of the documents and links between them. During the development hypertext has clear hierarchical structure, but a lot of cross-hyperlinks appeared which can destroy system of the hypertext so the system quality control can be lost during realization. The user can be confused and stop hypertext reading that can be extremely undesirable to the developers during using such system.

The first part of the work is devoted to the research of the hypertext system analysis experience and the role of the mathematical graph theory as the aim of the hypertext formalization.

The second part of the work is devoted to formalization of the problem of the losing control on the structure of the enlarging hypertext systems; criteria of the formal grade quality hypertext links are distinguished.

The example of the program, realizing developed method at T-SQL language and its application for real hypertext systems analysis is shown at the third part of the work.

As the accent is the necessity of the using of the iteration method adaptation and the expediency of the well-defined order criteria should be used during the hypertext optimization.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	7
1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ.....	11
1.1. Теория построения гипертекстовых систем.....	11
1.2. Теория графов и гипертекстовые системы.....	14
1.3. Определение основных понятий.....	15
1.4. Опыт анализа структуры информационных систем.....	16
2. РАЗРАБОТКА МЕТОДА АНАЛИЗА	19
2.1. Расчет характеристик графа гипертекста.....	19
2.1.1. Матрица расстояний	19
2.1.2. Преобразованная матрица расстояний	20
2.2. Постановка проблемы.....	21
2.3. Допущения.....	22
2.3.1. Отвлечение от содержимого документов	22
2.3.2. Простой граф гипертекстовой системы	23
2.3.3. Достижимость в разных множествах	23
2.4. Критерии оптимальности ссылочной структуры.....	24
2.4.1. Преобразованное расстояние графа.....	25
2.4.2. Доля отсутствующих путей	26
2.4.3. Индекс информационной компактности.....	26
2.5. Другие показатели качества гипертекста.....	27
2.5.1. Распределение значений в матрице расстояний.....	27
2.5.2. Мосты графа	28
2.5.3. Истоки и стоки	28
3. ОЦЕНКА РЕАЛЬНЫХ ГИПЕРТЕКСТОВЫХ СИСТЕМ.....	29
3.1. Сбор информации о гипертекстовых системах.....	29
3.2. Подготовка данных для анализа, их проверка и очистка.....	30
3.2.1. Перенос информации из MySQL в Microsoft SQL Server	30
3.2.2. Проверка целостности данных	32
3.2.3. Обзор очищенных данных	34
3.3. Результаты расчетов.....	35
3.4. Рекомендации по применению метода.....	36
ЗАКЛЮЧЕНИЕ.....	38
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	39
ПРИЛОЖЕНИЯ	41
Приложение I. SQL-программа расчета характеристик.....	41

ВВЕДЕНИЕ

В 1998-х году начался так называемый «Бум Интернет», продолжавшийся до начала 2000-х годов и завершившийся обвалом и банкротством ряда компаний электронной коммерции. После 2004 года наблюдается вторая волна этого бума, проявляющаяся в появлении все новых веб-сайтов, предоставляющих уже не просто услуги информации вроде электронных библиотек и каталогов, но широкий спектр бизнес услуг, в частности, Интернет-магазины и корпоративные веб-порталы [10]. Крупные компании интегрируют веб-сайты с внутренними хранилищами данных так, чтобы клиенты и партнеры имели актуальную информацию в любой момент времени. Характерной чертой современных веб-сайтов стало большое количество документов (номенклатура товаров в электронных магазинах содержит тысячи и десятки тысяч наименований) и внушительная пересеченность этого пространства документами ссылками. Однако обратной стороной этого высокого уровня предложения документов является затрудненность контроля качества и оптимальности построения ссылочных структур в гипертекстовых системах.

Зачем необходимо оценивать качество ссылочной структуры системы гипертекстовых документов (далее «СГД»)? В общей теории систем существует понятие «*свойство интегративности*», которое говорит о том, что система несводима к простой сумме её элементов. Говоря в контексте гипертекстовых систем, эффективность и информативность системы документов не равна простой сумме эффектов отдельных документов, структура ссылок также несет некоторую информацию и играет существенную роль в направлении прочтения документов [12]. Получается, что, в зависимости от наличия тех или иных связей между документами, гипертекстовая система может быть воспринята (прочтена) по-разному. Это можно сравнить с перестановками блоков текста в обычном линейном тексте, что может изменить смысловое содержание самым радикальным образом, так что внимательная организация структуры ссылок в многомерном гипертекстовом пространстве является необходимым условием для достижения цели гипертекста – *предоставления информации читателю*.

В данный момент гипертекстовые системы представлены главным образом множеством World Wide Web, в котором присутствует более 5 миллиардов документов, количество ссылок между ними, естественно, во много раз больше [15]. Также существуют отсоединенные от World Wide Web группы гипертекстовых документов, к примеру, внутренние корпоративные веб-сайты компаний и организаций. В данной работе сделана попытка дать рекомендации по анализу ссылочной структуры таких *небольших закрытых*

систем (по масштабам мирового пространства веб-документов), а также *выделенных по какому-либо признаку множеств* документов.

Цель данной работы – разработка *методической поддержки* разработчиков гипертекстовых систем, дающей рекомендации по построению ссылочных структур, удобных для просмотра пользователем. Задачи, решаемые в работе:

- 1) постановка задачи анализа ссылочной структуры;
- 2) выработка рекомендаций по сбору информации о гипертекстовых системах;
- 3) формулировка нескольких критериев качества ссылочной структуры;
- 4) апробация выработанных критериев на реальных гипертекстовых системах.

Гипертекст переводится буквально как «более чем текст». Сам принцип гипертекста является одним из лучших изобретений в области информатики, по словам изобретателя термина «гипертекст» Теда Нильсена, цель гипертекста – спасти мир [14]. Спасение подразумевается от моря информации, ориентироваться в котором стало сложно как никогда. Помощью в ориентации является нелинейный способ представления текста, позволяющий перемещаться между обозримыми блоками информации, без необходимости последовательного прочтения больших объемов, как это традиционно делается в печатных книгах. Технологически гипертекстовые документы размещаются на *веб-серверах* и просматриваются с помощью *веб-клиентов* (браузеров), которые обрабатывают ссылки, обеспечивая возможность перехода к другим документам [4].

В этой области уже проведена масса исследований, однако, подавляющая их часть проводится на Западе, пространство СНГ остается на периферии. К примеру, в 1999 году было проведено исследование в глобальных масштабах, в результате которого были выяснены характеристики ссылочной структуры всемирного множества гипертекстовых документов [15]. Существуют исследовательские работы по применению гипертекстовых систем в построении образовательных курсов, в которых указывается, что многомерное пространство гипертекстовых документов органично соответствует параллельным и многомерным мыслительным моделям человека, образуя, таким образом, технологию быстрого и эффективного изучения материала, особенно в дистанционном образовании [9]. Большинство исследований было проведено в 90-х годах прошлого столетия, однако широкого использования в работе фирм веб-разработки их результаты не получили.

Об **актуальности** тематики проводимого исследования говорит, в первую очередь, исследуемая область. Фактически это – весь современный Интернет, состоящий из миллионов веб-сайтов и миллиардов гипертекстовых документов, большая часть пользователей ассоциирует Интернет именно с веб-сайтами, поскольку с другими сервисами

или понятиями этой глобальной компьютерной сети просто не сталкиваются. В настоящее время почтовые, файловые и прочие сервисы предоставляются через интерфейс гипертекстовых страниц, собранных в веб-сайты. Масштабы современных гипертекстовых систем – это десятки тысяч документов и сотни тысяч ссылок, сохранить в этих пересеченных множествах удобство перемещения пользователя, не упустив из виду важных участков, становится практически невозможным без инструментальной и методологической поддержки.

Для определения **научной новизны** работы, прежде всего, был сделан поиск по доступным источникам, с целью найти материалы по данной тематике. Источники были рассмотрены в двух категориях:

- 1) печатные издания и книги;
- 2) электронные ресурсы, Интернет.

Печатные издания являются классическим полем поиска научных материалов, однако, современные электронные источники обладают рядом ключевых преимуществ, таких как полнотекстовый и тематический поиск, что весьма затруднено в собраниях твердых (бумажных) копий. Электронные источники особенно выгодны с точки зрения доступных объемов и скорости доступа к конечной информации, среди их недостатков можно отметить лишь недостаточную авторитетность информации, затрудненность проверки ее достоверности. С учетом всего вышесказанного, а также для поддержания полноты обзора поиск велся в обеих группах.

Основным результатом поиска в печатных изданиях являются упоминания о том, что пространство всемирной гипертекстовой паутины представляет собой граф (что существенно для нашей работы) и общие сведения о технологии построения гипертекстовых документов. Ближе всех к тематике исследования статья «Транзитивное замыкание» в журнале Windows IT Pro [1], рассматривающая построение транзитивного замыкания (показателя достижимости) и поиск кратчайших путей в графе, формализующем СГД. Однако дальнейшего анализа полученных данных о путях графа в статье не осуществляется.

Результаты поиска в Интернет более обширны и интересны, они подразделяются на две группы – русскоязычные и англоязычные источники. Среди русскоязычного Интернета большая часть результатов поиска по ключевым словам «оптимальность и оценка информационных структур» составлена бесполезными с нашей точки зрения статьями об оптимальности структур размещения финансовых капиталов, а также обещаниями студий веб-дизайна построить «сайт оптимальной структуры за умеренную плату». Самыми интересной среди русскоязычных источников оказались тезисы В.П. Полковникова на Конференции молодых ученых по математике, математическому моделированию и

информатике в Новосибирске в 2001 году, где сделана попытка сформулировать *показатель суммарной содержательной ценности* для СГД.

Значительно более продуктивным оказался поиск в англоязычной части Интернет, и это показывает, что на пространстве бывшего СССР вопросы анализа структуры гипертекстовой части Интернет проработаны крайне мало. В источниках на английском языке для систем гипертекстовых документов чаще используется краткий термин “web” (сеть, паутина, в русской транскрипции - «веб») и он будет использоваться наряду с «СГД» для краткости.

Результаты поиска на английском языке весьма разнообразны и большая их часть соответствует тематике исследования, однако даже просто прочтение всей массы материала заняло бы целые годы, поэтому исследование опирается на выборочные работы. Особо хочется отметить опубликованные на сайте всемирного консорциума Интернет материалы 9 международной конференции World Wide Web, среди которых есть данные *анализа глобальной ссылочной структуры* веб за 2000 год с использованием теории графов [15].

Таким образом, *данная работа не является кардинально новым* направлением исследований, однако акцент на практической применимости результатов исследований и разрабатываемых подходов, а также попытки сформулировать закономерности итоговой информационной структуры в зависимости от использования в ней различных приемов построения ссылок между документами, представляют некоторый новый взгляд на проблему, по крайней мере, на территории СНГ.

1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Теория построения гипертекстовых систем

История идеи связанных систем документов восходит к работам Ванневары Буша в 40-е годы 20 века, однако технологической основы для развития она в те времена не имела. Сам термин «гипертекст» появился позднее, в 1965 году, в работах Теда Нильсена. И все же эта идея ждала своего триумфального часа, когда появится удобный механизм просмотра и перемещения по документам [14].

Формально гипертекст определяется как представление текстовой информации в виде сети, в которой читатели получают свободу перемещаться нелинейным образом.

Даже после появления сети Интернет гипертекст не слишком быстро завоевывал популярность, так как не существовало удобных программных средств для использования всех его возможностей. Ситуация изменилась в 1993 году после выпуска программы Mosaic, с помощью которой просматривать и перемещаться между гипертекстовыми документами стало просто, для этого нужно было лишь кликнуть мышкой по гиперссылке. С 1994 года направляющую роль в глобальных гипертекстовых системах играет консорциум W3C, разрабатывающий стандарты и рекомендации для всемирной паутины World Wide Web [17].

Технологически WWW использует для передачи данных инфраструктуру глобальной компьютерной сети Интернет, и это является необходимым, но недостаточным условием полнофункциональной гипертекстовой системы. Основная суть технологии – архитектура клиент-сервер, с особым серверным и клиентским программным обеспечением [4]. Работа веб-технологии проходит в режиме запрос-ответ (Рис. 1).

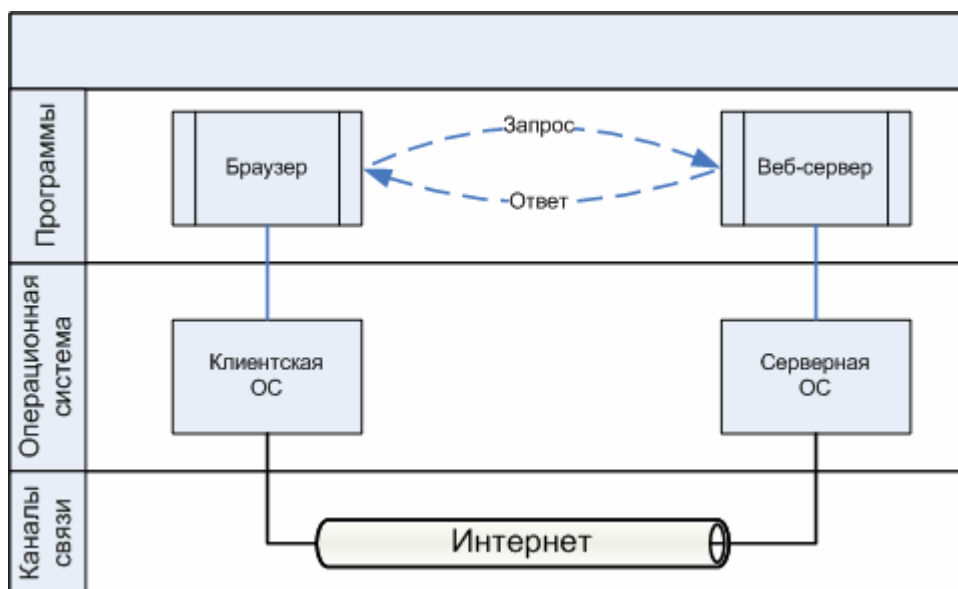


Рис. 1. Обобщенная архитектура World Wide Web

Операционные системы клиента и сервера обслуживают потребности приложений в связи посредством каналов Интернет, которые собственно передают данные от клиента к серверу и обратно.

Веб-сервер принимает от клиента запрос на документ и посылает ответ на этот запрос, в ответе содержится статус ошибки, или запрошенный документ. Веб-клиент (браузер) отображает информацию пользователю и предоставляет возможность выполнить следующий запрос на сервер (возможно отличный от предыдущего) посредством активации гиперссылки.

Крайне важное место в этой схеме занимает вопрос идентификации документов в множестве веб и адресация. В данный момент для идентификации веб-документов используются унифицированные указатели ресурса URL (Uniform Resource Locator), позволяющие направить запрос клиента по сети к однозначно определенному серверу и документу на нем. Любая гиперссылка – это объект, с которым ассоциирован URL, указывающий на веб-документ. Один документ может иметь множество ссылок на другие документы, равно как и сам быть предметом множества ссылок из других документов. Таким образом, документы и ссылки образуют сеть произвольной структуры. В этой сети могут быть близко связанные документы, относящиеся к одной группе URL, например, в одной зоне доменных имен DNS, а также ссылки на документы других серверов. Один из примеров системы связанных документов представлен на Рис. 2.

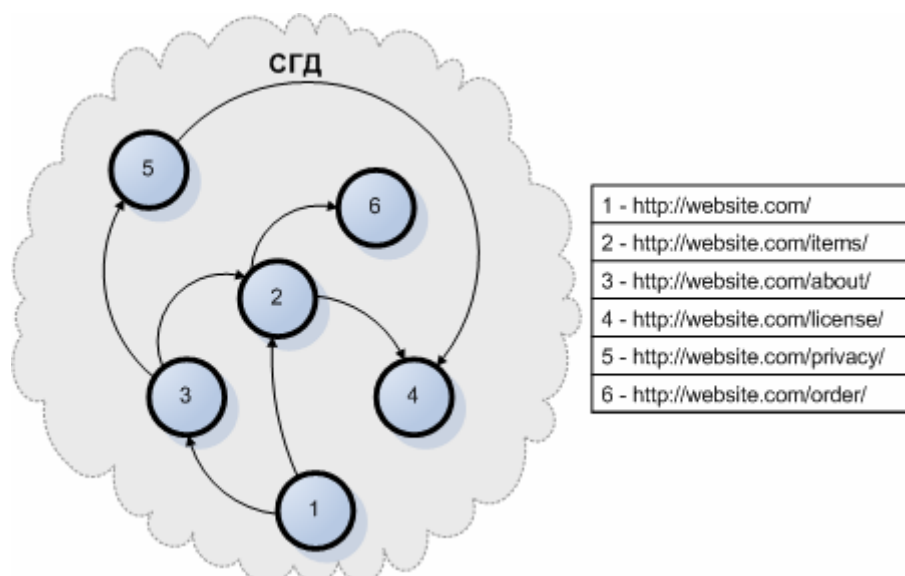


Рис. 2. Схема простой гипертекстовой системы, 6 документов и 7 ссылок

Следует отметить, что подобную схему имеет смысл строить лишь для небольших, порядка десятков документов, систем, так как графическое представление большой системы представляет собой крайне насыщенную узлами и соединениями схему. Немного улучшает ситуацию применение трехмерных моделей, но системы порядка сотен документов даже в трехмерной проекции крайне запутаны для восприятия, и это говорит о важности методов расчета характеристик гипертекстовых систем [2].

Если рассматривать общий случай, то веб-документом является любой файл, имеющий адрес URL и доступный в веб-системе, однако ограничимся гипертекстовыми документами, так как основными источниками информации и формирователями ссылочной структуры выступают именно они.

Современные гипертекстовые документы разрабатываются с использованием стандартов HTML и XHTML, описывающих язык разметки гипертекстовых документов. С помощью инструкций этого языка в документ включается текст, гиперссылки, а также множество других объектов, к примеру, графические изображения, формы ввода информации и программы на языках сценариев. Описание языков разметки и приемов веб-программирования находится за рамками данной работы и рассматриваться не будет.

Существенно, что современные программы веб-серверов позволяют строить веб-документы *динамически*. Когда веб только зарождался, документ представлял собой файл на диске, и задачей веб-сервера было лишь взять запрошенный файл и передать клиенту. Динамическое формирование веб-документа позволяет строить файл «на лету», используя для этого программу с разветвленной логикой, базы данных и практически неограниченный арсенал современных информационных технологий. Это приводит к тому, что два документа, запрошенные по одному URL-адресу, даже в течение короткого интервала

времени, могут существенно отличаться друг от друга. Таким образом, рассматривая современные множества гипертекстовых документов, можно говорить лишь о некотором срезе динамической веб-системы, отражающем ее состояние даже не на момент времени, а на совокупность моментов, в которые документы запрашивались, так называемый *обход системы (crawl)*. В этой ситуации документ идентифицируется не информационным содержимым, в виду его непостоянства, а фиксированным URL-адресом.

Если рассматривать большие масштабы гипертекстовых систем с динамическим формированием веб-документов, то они находятся в некотором динамическом равновесии своей ссылочной структуры, изменяясь незначительно, как бы колеблясь вокруг состояния покоя.

1.2. Теория графов и гипертекстовые системы

Большие системы любой природы состоят из множества элементов, между которыми имеются соединения. Для анализа структуры этих соединений используется математический аппарат теории графов, которая описывает объекты, состоящие из связанных элементов, а также методы определения отношений объектов в зависимости от структуры связей между ними. Основной вопрос, на который дает ответ теория графов, это *существование пути от одного узла графа к другому*.

Первым шагом в применении теории графов является **формализация** реальной системы в абстрактный объект – *граф*, чтобы изучить его характеристики и сделать на основании этого выводы о реальной системе, возможно, дать информацию для принятия решения об изменениях в системе.

В теории графов существует два типа графов – неориентированные (собственно графы) и ориентированные (орграфы), алгоритмы и характеристики для каждого типа немного отличаются. В зависимости от природы связей система формализуется либо в первый тип, либо во второй. Неориентированными графами описываются системы, для которых направление связи не имеет значения, к примеру, полнодуплексные каналы передачи данных, в которых сигнал распространяется в любом направлении, или тексты с перекрестными ссылками «вперед-назад». Орграфы описывают системы с направленными связями, к примеру, сети силового электрического питания, системы гипертекстовых документов и ссылок.

Граф является абстрактным понятием, один и тот же граф можно представить в разном виде:

- 1) в виде графической схемы (в двухмерной или трехмерной проекции), такое представление имеет смысл для графов порядка десятков, реже сотен узлов, когда нужна визуализация структуры связей;
- 2) в виде матрицы смежности, такое представление применяется для графов порядка десятков, реже сотен узлов, поскольку матрица занимает площадь (или объем памяти) пропорциональную квадрату количества узлов;
- 3) в виде списка смежных узлов, это представление позволяет сохранить граф в таблицах БД или списочных структурах, что удобно для обработки графов больших размерностей (тысячи узлов и более).

Какой бы способ не выбрать, граф будет одним и тем же и это дает свободу применения разных представлений, в зависимости от конкретной потребности.

Алгоритмы на графах делятся на две больших группы: поиск «в глубину» и поиск «в ширину». Первые используются для поиска, к примеру, мостов, когда нет требования поиска кратчайшего пути, а вторые – когда необходимо искать именно кратчайшие пути в графе.

1.3. Определение основных понятий

В результате продолжения изучения источников был составлен краткий глоссарий, содержащий определения понятий теории графов, используемые в области анализа информационных структур. Не ставится цель объять всю теорию графов, в которой присутствуют сотни понятий, а возьмем лишь касающиеся исследуемой области определения, которые будут использоваться далее в работе:

Формально **граф** G – это совокупность двух множеств (V, E) , в котором V – это **вершины**, а E – множество пар $e_i=(v_{i1}, v_{i2})$, $v_{ij} \in V$. Если порядок элементов пар e_i имеет значение, значит, имеем **ориентированный граф**, иначе **граф неориентированный** [6].

Узел – это вершина в орграфе. **Ребро** – это элемент множества E . В ориентированных графах ребра называются **дугами** [13].

Вершина графа v и некоторое его ребро e называются **инцидентными**, если $e=(v,w)$ или $e=(w,v)$, где w – некоторая вершина графа. Если вершины соединяет ребро, то они называются **смежными**.

Степень вершины – число инцидентных ей ребер. **Полустепень исхода** – количество дуг, исходящих из вершины. **Полустепень захода** – количество дуг, входящих в вершину.

Исток – узел, полустепень захода которого равна нулю. Аналогично **сток** – узел, полустепень исхода которого равна нулю.

Петля – это ребро, инцидентное единственной вершине, то есть $v_{i1}=v_{i2}$

Мультиребро – ребро, которое встречается в множестве E несколько раз.

Простой (обыкновенный) граф – граф без петель и мультиребер.

Маршрут или **путь** – чередующаяся последовательность $x_1, u_1, x_2, u_2, \dots, x_k$ вершин x_i и ребер u_i , обладающая тем свойством, что любая пара соседних элементов инцидентна.

Цепь – маршрут, в котором все ребра различны. **Простая цепь** – цепь, в которой все вершины различны. **Цикл** – цепь, в которой первая вершина идентична последней.

Если существует маршрут из вершины x в y , говорят что y **достижима** из x .

Матрица смежности графа – квадратная булева матрица, в которой $a_{ij}=1$, если существует ребро $e=(v_i, v_j)$ и $a_{ij}=0$, если такого ребра не существует. Для неориентированного графа матрица симметрична относительно главной диагонали. Главная диагональ всегда заполнена нулями.

Матрица расстояний (кратчайших цепей) графа G – квадратная матрица $D=\{d_{ij}\}$, в которой элемент $d_{ij}=r(x_i, x_j)$, т.е. численно равен расстоянию от вершины x_i до вершины x_j в графе G . Если из x_i недостижима вершина x_j , то $d_{ij}=\infty$. Главная диагональ всегда заполнена нулями.

Связный граф – неориентированный граф в котором из любой вершины можно найти цепь в любую другую вершину. Несвязный граф – распадается на **компоненты связности** (максимальные связные подграфы).

Мост – ребро графа, удаление которого увеличивает число его компонент связности.

Сильно-связный граф – орграф, в котором любая вершина достижима из любой вершины.

Полный граф – обыкновенный граф, в котором любая пара вершин смежна.

Эксцентриситет вершины – максимальное расстояние от v до других вершин графа.

Диаметр графа – максимальный эксцентриситет его вершин.

Средний диаметр – сумма кратчайших расстояний между всеми парами вершин графа, деленная на число пар вершин.

1.4. Опыт анализа структуры информационных систем

Из результатов изучения печатных источников и Интернет следует, что вопросы анализа структуры информационных систем стоят давно и неизменно остры в системах большого масштаба, независимо от характера и области. Существует очень много работ и приемов анализа организационной структуры предприятий, информационной структуры организаций и образований, рассматриваются также структуры размещения фондов и капиталов на различных рынках и их эффективность. Также встречаются работы по анализу внутренней структуры текстовых блоков, но они относятся ближе к области лингвистики и

психологии, а не информатики. Все изученные работы применяют для проведения анализа математический аппарат теории графов, поскольку он является адекватным инструментом формализации систем.

В контексте информационной структуры организации конечная цель анализа – получение информации для принятия решения об изменении этой структуры, ликвидации узких мест и повышение эффективности информационного обмена. Как увидим далее, для гипертекстовых систем существует достаточно близкая аналогия с этими целями. Однако масштабы гипертекстовых систем на порядки превышают информационные системы предприятия, и это лишает человека возможности легко принять решение после анализа схемы, иллюстрирующей информационную структуру. Действительно, количество объектов информационной структуры организации обычно ограничивается несколькими десятками, а количество документов в современном веб-сайте чаще всего превышает сотню и доходит до нескольких тысяч. В таких условиях говорить об обозримой графической схеме, таблице, или словесном описании невозможно.

Еще одним сегментом в анализе информационных структур является анализ исходных текстов программ, проводимый с целью адаптации последовательной программы для исполнения на машинах с параллельной архитектурой, класса суперЭВМ [7]. Масштабы информационных структур в этой сфере близки к гипертекстовым системам, однако цели и методы такого анализа лежат в совершенно иной плоскости.

И все же интересен анализ именно гипертекстовых систем, с учетом всех особенностей исследуемой среды. В этом сегменте в 1999 году группой ученых были проведены достаточно серьезные исследования, в результате которых были получены фактические данные о глобальной структуре гипертекстовой сети World Wide Web. По итогам исследования все множество документов и ссылок разделилось на 5 больших частей: сильно-связная компонента, истоки, стоки, «усики» и отсоединенные компоненты [15]. Однако ценность этих результатов для владельца или разработчика отдельного веб-сайта или гипертекстовой системы (что не есть одно и то же) чисто информативная и общепознавательная. Для разработчика гораздо важнее иметь под рукой *четкие критерии оценки* своей системы, а еще лучше *прикладной инструмент*, делающий этот анализ автоматически и предоставляющий готовую информацию для принятия решений об изменениях в структуре.

Еще одна интересная попытка вывести критерий оценки гипертекстовой системы была сделана в 2000 году в Новосибирске. На Конференции молодых ученых по математике, математическому моделированию и информатике Полковников Е.В. сформулировал

показатель суммарной содержательной ценности информационных блоков в системах гипертекстовых документов, как

$$w^i = w(v_s) + \sum_{v_i \in V^i v_s} \frac{w(v_i)}{\sum_{e_j \in (v_s, v_i)} (1 + p(e_j))} \quad (1)$$

где v_s – начальная вершина просмотра, $w(v_i)$ – информационная ценность блока, $p(e_j)$ – стоимость перехода по ссылке [11]. Однако у такого подхода есть ряд недостатков. Во-первых, содержательная ценность информационного блока является весьма субъективной величиной, зависящей от конкретного читателя, любая попытка сказать, что один массив информации ценнее другого сродни проявлению шовинизма и неприемлема. Во-вторых, понятие «стоимости перехода» в современном гипертекстовом пространстве не имеет смысла, поскольку зависит от совершенно неконтролируемого набора факторов, таких как состояние сетей коммуникаций, клиентской и серверной машин, оформления гиперссылок и все той же субъективной ценности ссылки для конкретного читателя; представляется сомнительным наличие возможности оценивать эту стоимость по единой методике.

Очень интересна работа Г.Е. Кедровой «Методы оптимизации компьютерной обучающей среды по лингвистике для систем дистанционного обучения в Интернете», где приведен пример построения гипертекстовой системы с анализом структуры связей в ней и учетом вопросов оптимальности этой структуры исходя из психологии читателя. В качестве критерия оптимальности ссылочной структуры работа опирается на индекс информационной компактности [9], который является одним из ключевых в оценке оптимальности гипертекстовой системы. Индекс информационной компактности подробно рассмотрен в п. 2.4.2.

2. РАЗРАБОТКА МЕТОДА АНАЛИЗА

2.1. Расчет характеристик графа гипертекста

Итак, твердо известно, что гипертекстовая система формализуется в математический ориентированный граф. Прежде чем приступить к формулированию критериев оптимальности гипертекста, определим, какие показатели и характеристики исходно доступны для расчета этом графе.

2.1.1. Матрица расстояний

Первым этапом на пути к определению характеристик графа является расчет в нем кратчайших путей, дающий матрицу расстояний орграфа. Отправной точкой в этих расчетах является матрица смежности, содержащая информацию о прямых ссылках в гипертексте, вопросы получения информации об этих ссылках рассмотрены ниже, в 3 главе.

Матрица смежности квадратная, ее строки представляют документы, *из* которых идут ссылки, а столбцы – *куда* они идут. Если прямая ссылка между документами существует, то соответствующая позиция матрицы равна единице, а если нет – то нулю. Таким образом, матрица смежности *булева*. По главной диагонали матрицы смежности всегда стоят нули [13]. На Рис. 3 приведен пример матрицы смежности, соответствующий схеме СГД из подраздела 1.1.

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	0	0	0	1	0	1
3	0	1	0	0	1	0
4	0	0	0	0	0	0
5	0	0	0	1	0	0
6	0	0	0	0	0	0

Рис. 3. Матрица смежности орграфа

Количество позиций в матрице смежности растет пропорционально квадрату количества узлов n и это существенно для определения затрат на вычисления в этой матрице, которые возрастают резко с ростом множества документов.

Далее происходит нахождение всех возможных путей в графе волновым алгоритмом [8], и нулевые значения заполняются числами, соответствующими длине кратчайшего пути между узлами. По главной диагонали матрицы остаются нули, показывая наличие бесконечно малого пути перехода от документа к нему же самому. При отсутствии пути через промежуточные узлы в матрице ставится значение ∞ , показывающее невозможность даже опосредованного перехода из одного документа в другой (Рис. 4).

	1	2	3	4	5	6
1	0	1	1	2	2	2
2	∞	0	∞	1	∞	1
3	∞	1	0	2	1	2
4	∞	∞	∞	0	∞	∞
5	∞	∞	∞	1	0	∞
6	∞	∞	∞	∞	∞	0

Рис. 4. Матрица расстояний орграфа

Сам **волновой алгоритм** является простым и очень надежным итерационным способом вычисления кратчайших путей в графе. На каждой итерации k алгоритма вычисляются пути длиной $k+1$ и вычисления продолжают до тех пор, пока в ходе итерации обнаруживаются новые пути.

Исходно для вычислений по волновому алгоритму берется матрица смежности, и ее позиции, содержащие нули, кроме главной диагонали, очищаются, $e_{ij} = \emptyset$.

В процессе итерации k просматриваются все пустые позиции матрицы смежности $e_{ij} = \emptyset$, каждая из которых проверяется на существование пути e_{nj} из v_i в v_j через какой-либо косвенный узел графа e_{in} , имеющий прямую связь до v_j , то есть

$$e_{ij} = \begin{cases} k+1 & \text{если } e_{nj} \neq \emptyset \wedge e_{in} = 1 \\ \emptyset & \text{иначе} \end{cases} \quad (2)$$

Когда на очередной итерации новых путей в матрице не обнаружено, оставшиеся пустые позиции заполняются значением «бесконечность», поскольку таких путей не существует.

2.1.2. Преобразованная матрица расстояний

Недостатком матрицы расстояний является то, что оперирование с бесконечными величинами крайне неудобно и приводит к невозможности вычисления характеристик графа. К примеру, сумма расстояний графа, а также его диаметр равны бесконечности, что исключает возможность сравнения и оценки графов, в которых отсутствует хотя бы один возможный путь.

Эта проблема решается введением *преобразованной матрицы расстояний*, в которой позиции недостижимых узлов вместо бесконечностей заполняются константой преобразования K [5].

Константа преобразования K может выбираться в зависимости от целей преобразования, но с нашей точки зрения будет достаточно брать ее на единицу большей максимально возможной длины пути в графе данной размерности. Такой подход дает для недостижимых путей длину большую даже максимально возможной, но все же конечную.

Плюс к этому константа растет вместе с размерностями графа и не является постулированной величиной, как это было бы в случае фиксации ее на некотором заведомо «большом» значении. В нашем примере константа преобразования равна шести (Рис. 5).

	1	2	3	4	5	6
1	0	1	1	2	2	2
2	6	0	6	1	6	1
3	6	1	0	2	1	2
4	6	6	6	0	6	6
5	6	6	6	1	0	6
6	6	6	6	6	6	0

Рис. 5. Преобразованная матрица расстояний

Полученная в итоге преобразованная матрица расстояний готова к расчетам *конечных* численных характеристик ориентированного графа СГД.

2.2. Постановка проблемы

При первоначальном проектировании информационной структуры веб-сайта производится разбивка текста на блоки, между которыми проектируются четкие структурные связи, отражающие иерархию информации и предполагаемые направления ее прочтения (Рис. 6).

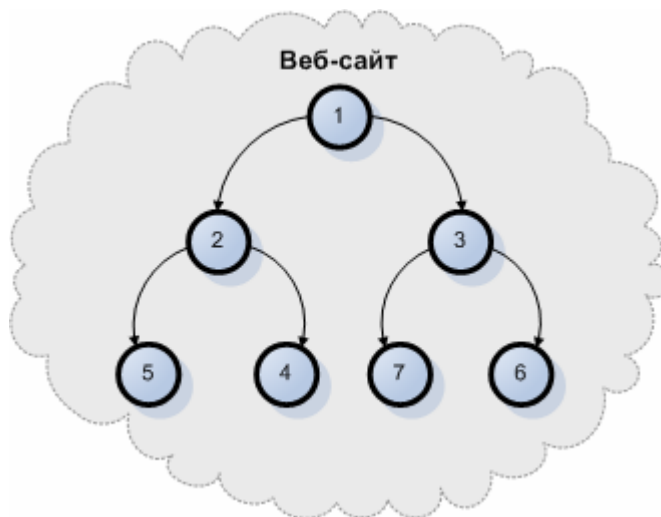


Рис. 6. Концептуальная схема веб-сайта, пример

Однако после перехода к проектированию и реализации гипертекста в системе появляется множество *перекрестных ссылок*. К примеру, организуются ссылки на стартовую страницу, ссылки из одних блоков текста в другие на одном уровне иерархии, карты сайта и алфавитные указатели (Рис. 7).

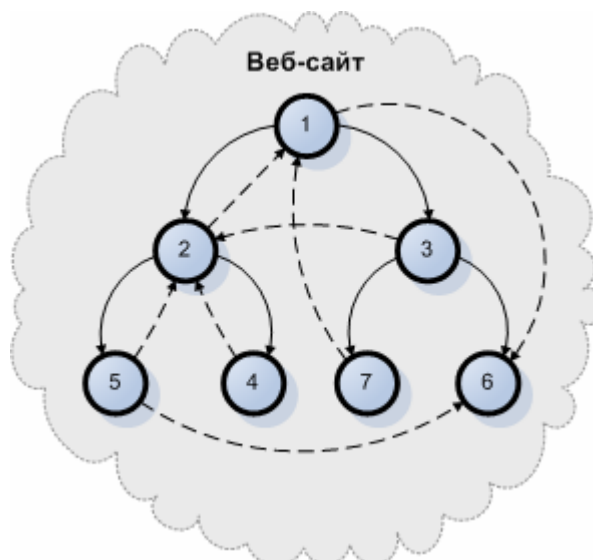


Рис. 7. Реальная ссылочная структура сайта, пунктиром указаны перекрестные ссылки

Для гипертекста это естественный процесс, в этом его предназначение – превращать линейный текст в многомерное пространство с множеством связей. Однако при увеличении масштабов системы до сотен и тысяч документов, и особенно при динамическом формировании страниц, структура ссылок выходит из-под контроля и перестает быть обозримой. Построить графическую схему, чтобы увидеть слабые места ссылочной структуры, становится нереально. В этой ситуации необходимо иметь критерии качества ссылочной структуры гипертекстовой системы, расчет которых возможно автоматизировать, а сами они должны быть формальны.

2.3. Допущения

2.3.1. Отвлечение от содержимого документов

Для реального читателя гипертекста каждый документ имеет большее или меньшее значение, прежде всего, в зависимости от целей самого читателя, его знаний и субъективных суждений. Также гиперссылки реально ощущаются читателем по-разному, одни более информативны, другие менее. Это зависит от множества факторов оформления, психологии, порой даже физиологии читателя (простейший пример – острота зрения читателя, которая влияет на восприятие мелкого текста). Таким образом, всесторонняя оценка гипертекста – это совокупность формальных и субъективных характеристик, адекватная лишь для конкретного наблюдателя.

Ввиду вышеизложенного невозможно иметь формальный метод *всесторонней* оценки и сравнения гипертекстовых систем, необходимы некоторые допущения, открывающие путь к формализации гипертекста. Одно из главных допущений в нашем подходе – это равноценность всех документов и ссылок. Таким образом, *анализируется не смысловое содержание гипертекста, а только его ссылочная структура.*

После принятия этого допущения гипертекстовая система прекрасно формализуется в невзвешенный ориентированный граф, без чего невозможна выработка формальных критериев оценки оптимальности и дальнейшее применение методов к реальным СГД.

2.3.2. Простой граф гипертекстовой системы

Граф гипертекстовой системы содержит в себе *петли*, так как в документе могут присутствовать ссылки на якоря в этом же документе, однако они никак не влияют на достижимость других документов.

Также в системе присутствуют не только единичные прямые ссылки, но и множественные, к примеру, организуются две строки навигации – верху и внизу, которые удваивают количество прямых связей между документами. В графе это отражается в виде *мультиребер*, которые также не оказывают влияние на достижимость документов.

Тем не менее, петли и мультиребра оказывают свое влияние на перемещение пользователя по гипертексту, однако в данный момент не разработано подхода к учету таких ссылок в критериях оценки СГД и они рассматриваться не будут, это второе допущение – *граф ссылочной структуры гипертекстовой системы является простым*.

2.3.3. Достижимость в разных множествах

Очень важным моментом является выбор и ограничение множества документов, которое будет анализироваться. Дело в том, что два документа, которые не имеют пути друг к другу в рамках одного веб-сайта, могут быть достижимы через внешние сайты (Рис. 8). Путь этот может быть неочевидным и запутанным, но все же его существование не исключено. Поэтому еще одно допущение, которое принимается, заключается в том, что *достижимость в надмножестве исследуемого множества документов не рассматривается*. Как правило, границы множества документов являются границами веб-сайта, и внутри этого множества ссылочная структура должна быть достаточной для достижения любого документа, разработчику СГД рассчитывать на то, что другие сайты предоставят ссылки для опосредованного достижения *недопустимо*.

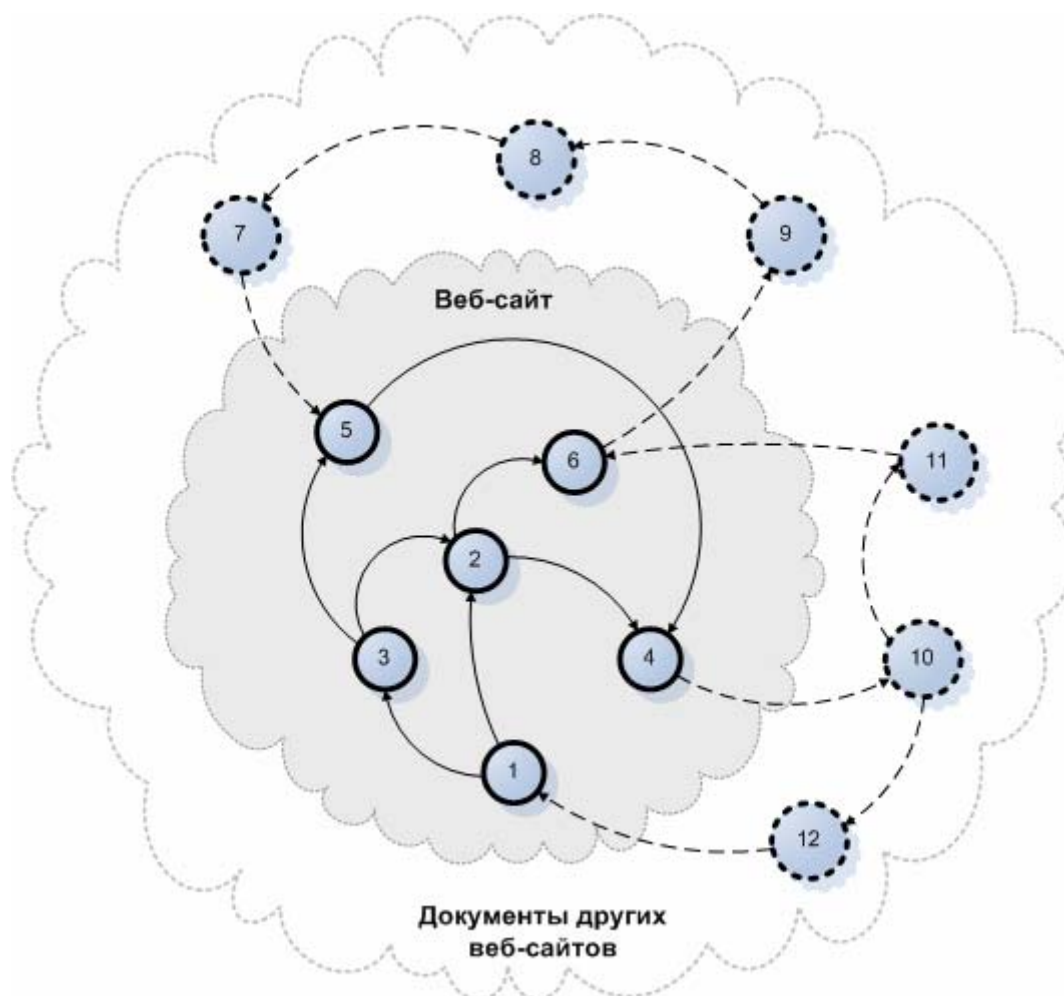


Рис. 8. Достижимость в разных множествах

2.4. Критерии оптимальности ссылочной структуры

Выяснилось, что гипертекстовая система представляет собой граф [4]. Зададимся вопросом – какие выводы позволяет сделать о гипертекстовой системе анализ графа? Какие критерии отражают оптимальность ссылочной структуры гипертекста?

В вопросе оптимальности гипертекстовой системы существенным является то, что хорошие характеристики математического графа *не дают гарантии оптимальности* реальной системы, ввиду приведенных выше допущений. Однако ожидается, что оценка оптимальности гипертекстовой системы по любому критерию будет коррелировать с субъективной оценкой наблюдателя-человека.

Необходимо понимать, что единого критерия оптимальности ссылок в СГД быть не может, критериев несколько, и они порой противоречат друг другу. Простейший пример – это индекс информационной компактности (рассмотрен ниже), который максимален для сильно-связного графа, однако этот граф отражает систему без структуризации информации, что равносильно полной бессистемности гипертекста и приводит к дезориентации читателя. Очевидно, что оптимизация гипертекста должна вестись по нескольким критериям.

Рассмотрим и охарактеризуем некоторые отдельные критерии оптимальности гипертекста, а затем опробуем их на реальных системах.

Существуют некоторые требования и пожелания к критериям, которые должны соблюдаться, чтобы достигнуть поставленной нами цели. Прежде всего, критерии должны быть формальны (не субъективны), иметь ясный метод расчета. Желательно, чтобы критерии предоставляли возможность сравнения разных гипертекстов между собой

2.4.1. Преобразованное расстояние графа

Первым интуитивно понятным критерием является длина путей в графе. Каждый читатель гипертекста понимает, что чем больше ссылок и промежуточных документов ему приходится пройти в поисках интересующего его блока информации, тем хуже он оценит гипертекстовую систему. Как только путь становится субъективно «слишком длинным», в зависимости от терпеливости читателя, человек ищет другой способ найти информацию, к примеру, пользуется полнотекстовым поиском [2].

Развитие полнотекстовых поисковых машин как средств глобального поиска информации в веб-среде привело к появлению понятия *поисковой оптимизации* (Search Engine Optimization, SEO), суть которой состоит в изменении гипертекста таким образом, чтобы ссылки на него как можно чаще появлялись в результатах поисковых машин. Приведем одно из высказываний фирмы, предоставляющей услуги поисковой оптимизации [16]:

«Внутренняя ссылочная структура, согласно SEO, общеизвестна как навигация web-сайта, является основным требованием для удобства посетителей и роботов поисковых систем. Логическая и интуитивная навигация помогает посетителю легко перемещаться по сайту и увеличивает степень отдачи (будь то в продажах, телефонных звонках или просто читателях). Хорошо составленная ссылочная структура помогает «паукам» поисковых систем с легкостью индексировать страницы сайта. В конечном счете, это облегчает пользователям поисковых машин задачу поиска необходимой информации».

Показателем длины путей в графе является преобразованное расстояние CD , равное сумме всех позиций в преобразованной матрице расстояний:

$$CD = \sum_i \sum_j C_{ij} \quad (3)$$

Наименьшее преобразованное расстояние имеет полный оргграф:

$$CD_{\min} = n^2 - n \quad (4)$$

Наибольшее преобразованное расстояние имеет полностью несвязный граф, не имеющий ни одной связи:

$$CD_{\max} = n \cdot (n^2 - n) \quad (5)$$

Недостатком этого критерия является то, что сравнить два разных гипертекста с его помощью можно, только если в них одинаковое количество документов. Однако гипертекст с количеством документов 100 и $CD=15\ 000$ может на практике иметь пути длиннее и быть хуже, чем гипертекст размером 200 узлов и $CD=40\ 000$. В нашем примере из подраздела 1.1 система имеет $CD=125$.

2.4.2. Доля отсутствующих путей

Хуже длинных путей лишь их полное отсутствие. Вторым понятным критерием является количество отсутствующих путей Q_m в графе. Максимальное количество отсутствующих путей равно n^2-n , минимальное – ноль. Однако это очень узкий критерий, использующийся для первых оценок СГД и выявления самых грубых ошибок в ссылочной структуре гипертекста.

К тому же, сравнение по такому критерию невозможно, так как гипертекстовые системы варьируются в своих размерах, и отсутствие десятка путей в СГД с тысячами документов не является катастрофой, тогда как в системе с пятнадцатью это весьма ощутимо. Поэтому в качестве критерия по отсутствию путей возьмем *долю отсутствующих путей*:

$$K_m = \frac{Q_m}{n^2 - n} \quad (6)$$

Значение доли отсутствующих путей колеблется в пределах [0;1] и допускает сравнение систем гипертекстовых документов между собой. В нашем примере доля отсутствующих путей $K_m=60\%$, это очень большое значение, показывающее плохую достижимость между документами в гипертексте.

Для расчета коэффициента достаточно матрицы транзитивного замыкания графа [1], однако, поскольку нам все равно придется строить матрицу расстояний, мы воспользуемся ею.

2.4.3. Индекс информационной компактности

В качестве одного из критериев оптимальности гипертекстовой системы предлагается использовать индекс компактности, отражающий пересеченность гипертекстовой системы ссылками и близость графа системы к полному. Индекс компактности описан в работе

Ботафого «Структурный анализ гипертекста. Выявление иерархий и полезные метрики» [5] и выражается как

$$C_p = \frac{CD_{\max} - CD}{CD_{\max} - CD_{\min}} \quad (7)$$

Поскольку константа преобразования принята равной количеству узлов в графе n , то после некоторых преобразований имеем

$$C_p = \frac{n^3 - n^2 - \sum_i \sum_j C_{ij}}{n^3 - 2n^2 + n} \quad (8)$$

Индекс компактности опирается на понятие *преобразованного расстояния* CD , но в отличие от него обладает замечательным свойством – он изменяется в пределах $[0;1]$ для любого графа, давая, таким образом, возможность сравнивать различные гипертексты между собой. Индекс компактности нашего примера $C_p=0,3667$.

Безусловно, индекс компактности является хорошим показателем связности гипертекста, однако не стоит забывать, что понятия эффективности и оптимальности существенно отличаются друг от друга и системы с высоким индексом приближаются к неструктурированным полным графам.

2.5. Другие показатели качества гипертекста

Ввиду отсутствия однозначного критерия оценки ссылочной структуры необходимо предоставлять лицу, принимающему решение о построении структуры гипертекста или изменениях в ней, дополнительную информацию о системе и узких местах в ней.

2.5.1. Распределение значений в матрице расстояний

Все перечисленные критерии являются показателем качества СГД, однако, их недостаточно для определения дальнейших действий разработчика гипертекста, после их расчета и получения неудовлетворительных результатов необходимо обнаружить узкие места системы. Для этого предлагается прежде всего изучить распределение длин путей в матрице расстояний, чтобы выяснить, на какой группе путей сосредоточить усилия оптимизации. Для нашего (весьма простого) примера распределение выглядит следующим образом:

- 1) 7 прямых ссылок;
- 2) 5 ссылок длиной 2;
- 3) 18 отсутствующих путей.

По этому распределению сразу видно, что проблему представляют отсутствующие пути в системе, а пути длиной 2 перехода немногочисленны и вполне допустимы.

2.5.2. Мосты графа

Также предлагается уделять больше внимания таким слабым местам гипертекстовой системы, какими являются единственные дуги между связными компонентами графа, иначе называемые *мостами*. Мосты в графе отражают единственные ссылки в гипертексте, по которым можно перейти из одной части гипертекста в другую. Наличие в гипертексте мостов означает, что читателю в каких-то случаях приходится проходить один и тот же путь, состоящий из двух документов и ссылки между ними. В такой ситуации у пользователя может возникнуть негативная реакция и ощущение, что он ходит по кругу, поэтому появление мостов в гипертексте является нежелательным.

Поиск мостов в графе является задачей, отличающейся по подходу от задачи поиска кратчайших путей и матрица расстояний для этого не пригодна. Подробно алгоритм поиска мостов рассмотрен в работе [3].

2.5.3. Истоки и стоки

Примечательными точками в гипертекстовой навигации являются стоки и истоки в терминологии теории графов (см. подразд. 1.3). *Сток* – это страница, на которую указывают ссылки, но сама она не позволяет никуда переместиться, так как ссылок не имеет. Это явный тупик в навигации, так как пользователю из этой точки необходимо выбираться посторонними средствами, например, вручную введя новый адрес в веб-клиенте. Необходимо избегать появления стоков в гипертексте, всегда предоставляя ссылки для выхода на другие документы. Сток легко определяется в матрице смежности как строка, заполненная нулями.

Исток – это антипод стока, документ, из которого можно переместиться к другим, но на него нет ссылки ни в одном документе этого множества. Попасть на такую страницу можно только по ссылке из надмножества, или вручную введя адрес документа. Очевидно, что истоки также нежелательны в гипертексте, навигация с ними *не достаточна*. Исток определяется в матрице смежности как столбец, заполненный нулями.

3. ОЦЕНКА РЕАЛЬНЫХ ГИПЕРТЕКСТОВЫХ СИСТЕМ

3.1. Сбор информации о гипертекстовых системах

Чтобы собрать информацию о структуре реальной гипертекстовой системы экспериментальным путем, необходимо пройти по всем ее ссылкам, достигнуть всех ее документов. Как правило, именно таким путем пользуются поисковые системы и некоторые утилиты анализа гипертекста. Поскольку информационные хранилища крупных поисковых систем являются коммерческой собственностью поддерживающих их компаний, получить в свое распоряжение информацию о широком множестве гипертекстовых документов Интернет не представляется возможным. Утилиты анализа ссылочной структуры гипертекста малочисленны и информация, получаемая от них, требует больших затрат на конверсию в удобный для анализа формат.

Благодаря тому, что на КАФ-Интернет функционирует информационно-поисковая система, удалось получить копию ее базы данных в учебных целях. Исходная БД содержала информацию о 8699 веб-серверах, 4989 страницах, 13002 ссылках. Структура БД была проанализирована и построена ее ER-диаграмма (Рис. 9).

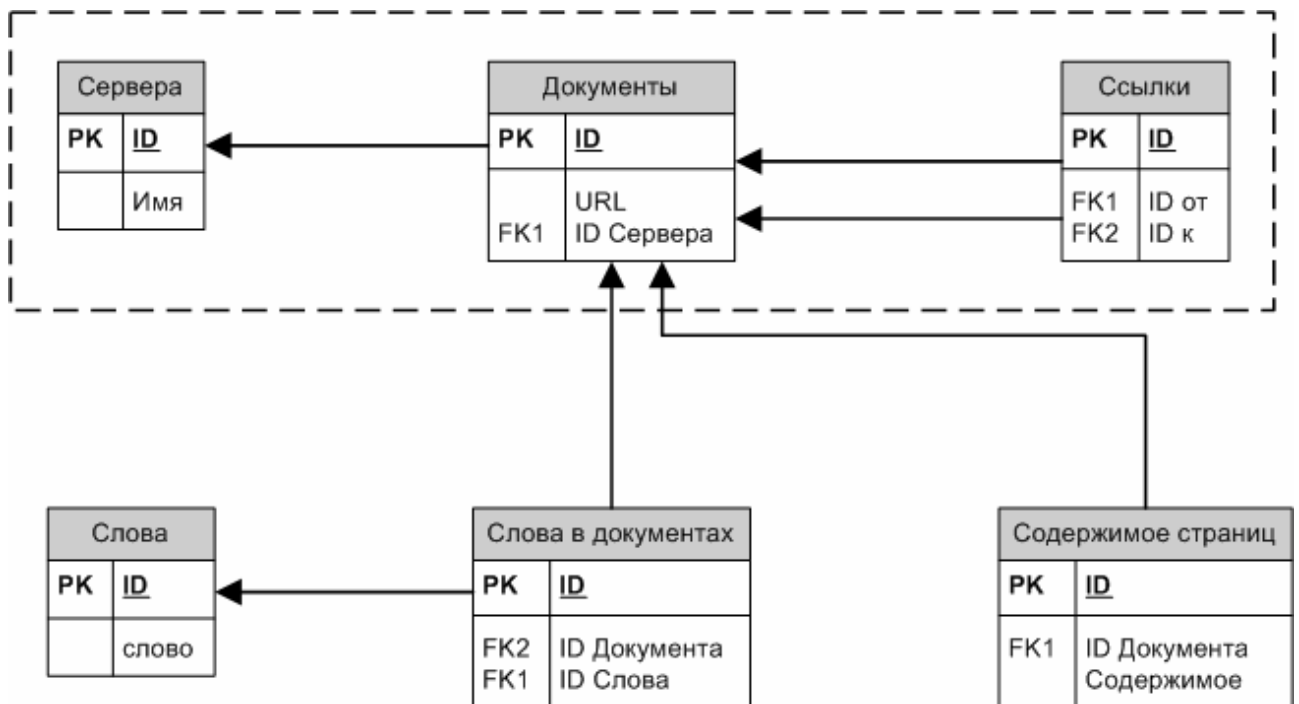


Рис. 9. ER-модель исходной БД

Из всех таблиц поисковой БД интересны лишь три, отмеченные на рисунке пунктиром, в них содержится информация, достаточная для анализа ссылочной структуры гипертекстовых документов. Таблицу *серверов* будем использовать для выделения множеств

документов и оценки отдельных веб-сайтов, *документы* и *ссылки* дадут собственно информацию о графах СГД.

3.2. Подготовка данных для анализа, их проверка и очистка

3.2.1. Перенос информации из MySQL в Microsoft SQL Server

Исходная база данных поисковой системы была получена в формате СУБД MySQL версии 4.2, так как поисковая машина КАФ-Интернет пользуется ей для хранения данных. Однако этот формат не подходит для наших исследований, поскольку представляет весьма ограниченный набор аналитических возможностей и построения сложных запросов.

Программный продукт Microsoft SQL Server 2000 обладает более широкими возможностями, позволяя писать программы на языке SQL, используя циклы и сложные объединения, поэтому база данных была конвертирована в формат Microsoft SQL Server с помощью специальной утилиты DTS (Рис. 10).

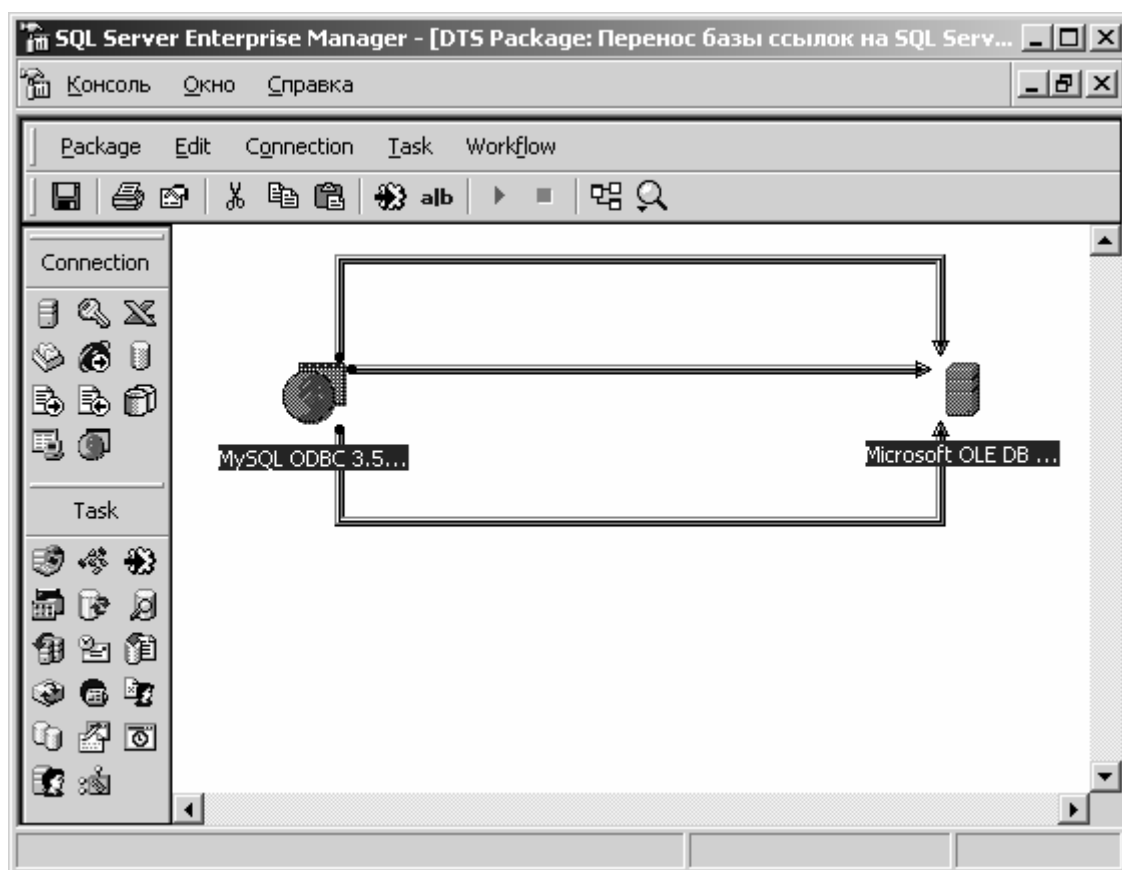


Рис. 10. Рабочее окно утилиты DTS

Процесс построения пакета DTS является простым, не будем раскрывать все его детали в данной работе, они несущественны для исследования. Итогом работы пакета явился успешный перенос данных из трех таблиц MySQL в соответствующие таблицы БД Microsoft SQL Server (Рис. 11).

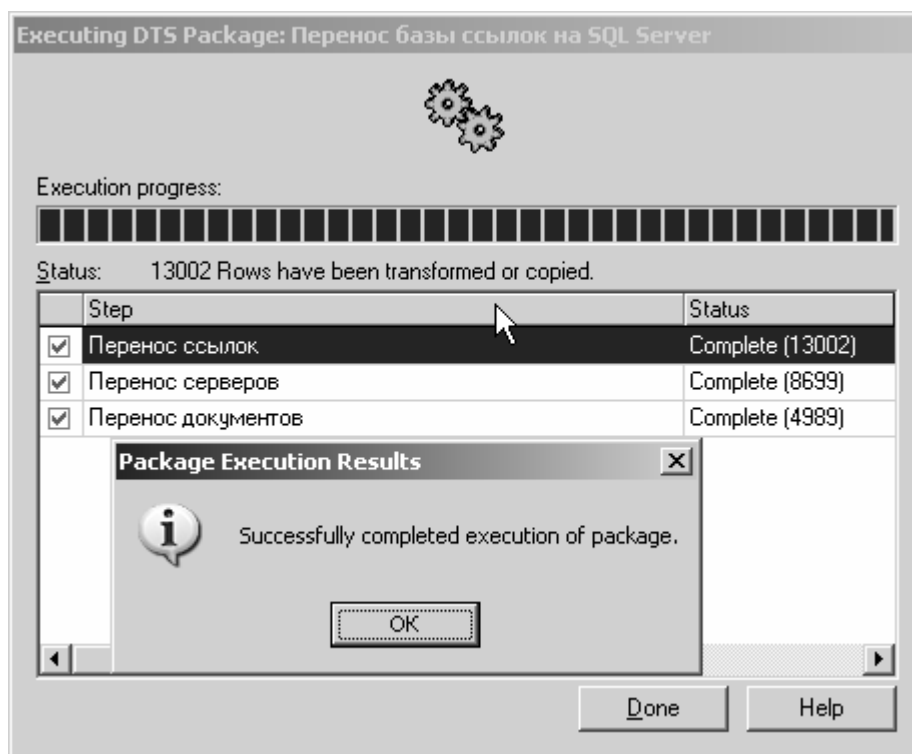


Рис. 11. Отчет об успешном завершении работы DTS

Следует сказать еще несколько слов о выборе инструментального средства проведения расчетов. Дело в том, что расчеты подобного рода не принято делать на языке SQL, так как реляционные БД для этого не предназначены. Расчеты на графах выполняются намного быстрее, если программа написана на языке Си или подобном ему компилируемом языке высокого уровня [3]. Тем не менее, воспользуемся именно SQL по нескольким причинам, прежде всего, по причине скорости написания и отладки программ, расплачиваясь за это лишь длительностью расчетов. Положительной стороной использования языка запросов является то, что СУБД не будет использовать для расчетов такие большие объемы оперативной памяти, как прикладная программа, и это снижает риск сбоев по причине заполнения ОЗУ. Поскольку в задачах проводимой работы не значит поиск эффективного инструментального средства анализа, выбирается удобство и гибкость в ущерб скорости расчетов, однако то, что для прикладного использования метода нужен эффективный инструмент расчетов, является *неоспоримым фактом*.

Графической утилитой для работы с SQL Server является SQL Query Analyzer, инструментальное средство редактирования и исполнения запросов на языке T-SQL (Рис. 12).

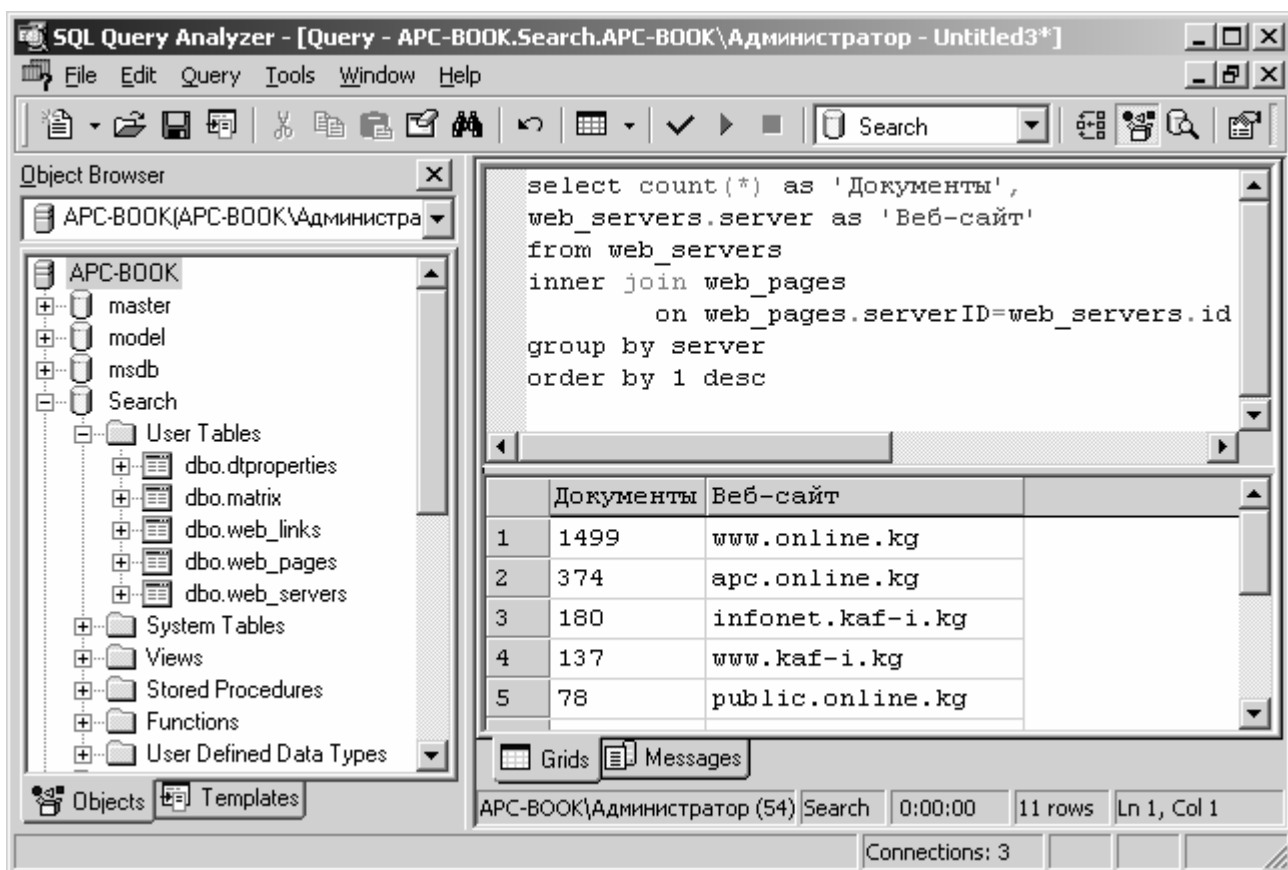


Рис. 12. Утилита SQL Query Analyzer

3.2.2. Проверка целостности данных

Прежде чем приступить к расчетам, необходимо убедиться, что исходные данные не содержат противоречий и ненужного «мусора» в виде веб-серверов, не имеющих относящихся к ним документов. Для начала нужно проверить, есть ли в БД записи о страницах, нарушающие связь один-ко-многим, которым не соответствуют сервера с помощью следующего SQL-запроса:

```
select URL, Web_servers.ID from web_pages
left join WEB_SERVERS on
WEB_SERVERS.ID=WEB_PAGES.ServerID
where WEB_SERVERS.ID is null
```

В результате было обнаружено 188 страниц без соответствующего сервера, все они, судя по URL, принадлежали известным серверам, но почему-то были неверно занесены в БД. Некорректные записи были исправлены SQL-запросом:

```

update WEB_PAGES
SET ServerID=H2.ID
from web_pages
  left join WEB_SERVERS on
WEB_SERVERS.ID=WEB_PAGES.ServerID
  inner join WEB_SERVERS H2 on left(right(URL, Len(URL)-
7), charindex('/', right(URL, Len(URL)-7))-1)=H2.server
where WEB_SERVERS.ID is null

```

Повторная проверка целостности по страницам показала, что ситуация исправлена успешно. Далее было проверено, есть ли страницы без инцидентных им ссылок – это технологически еще необработанные поисковой машиной документы:

```

select *
from web_pages
  left join web_links on IDfrom=Id or Idto=id
where IDfrom is null
order by Blcnt desc, flcnt desc

```

Таких документов нашлось 2715, они были удалены из базы данных запросом:

```

delete web_pages
from web_pages
  left join web_links on IDfrom=Id or Idto=id
where IDfrom is null

```

Следующим шагом проверено наличие в списке серверов записей, для которых нет ни одной страницы:

```

select WEB_SERVERS.* from WEB_SERVERS
  left join WEB_PAGES on
WEB_SERVERS.ID=WEB_PAGES.ServerID
where WEB_PAGES.ID is null

```

Обнаружилось 8688 таких записей, они были удалены командой:

```

delete from WEB_SERVERS where ID in (
select distinct WEB_SERVERS.ID
from WEB_SERVERS
  left join WEB_PAGES on
WEB_SERVERS.ID=WEB_PAGES.ServerID
where WEB_PAGES.ID is null)

```

Далее проверена целостность связи по ссылкам:

```

select WEB_LINKS.* from WEB_LINKS
  left join WEB_PAGES on
WEB_LINKS.IDfrom=WEB_PAGES.ID OR
WEB_LINKS.IDto=WEB_PAGES.ID
where web_pages.ID is null

```

Итог – все ссылки принадлежат существующим страницам, целостность в норме.

Последним шагом проверено наличие ссылок страниц самих на себя (петель), которые недопустимы для расчетов на простом графе (см. пункт 2.3.2):

```
select * from web_links where IDfrom=IDto
```

Обнаружилась 281 петля, и мы удалили их запросом:

```
delete from web_links where IDfrom=IDto
```

3.2.3. Обзор очищенных данных

В итоге очистки и подготовки данных получена БД в Microsoft SQL Server, описывающая 11 веб-серверов, 2274 страниц, 12721 ссылки. Поскольку запланировано проанализировать характеристики каждого веб-сайта, интересно количество документов, относящееся к каждому из них. Распределение документов по сайтам рассчитано группирующим SQL-запросом и графически представлено на Рис. 13.

```
select count(*), web_servers.server from web_servers
inner join web_pages on
web_pages.serverID=web_servers.id
group by server
order by 1 desc
```

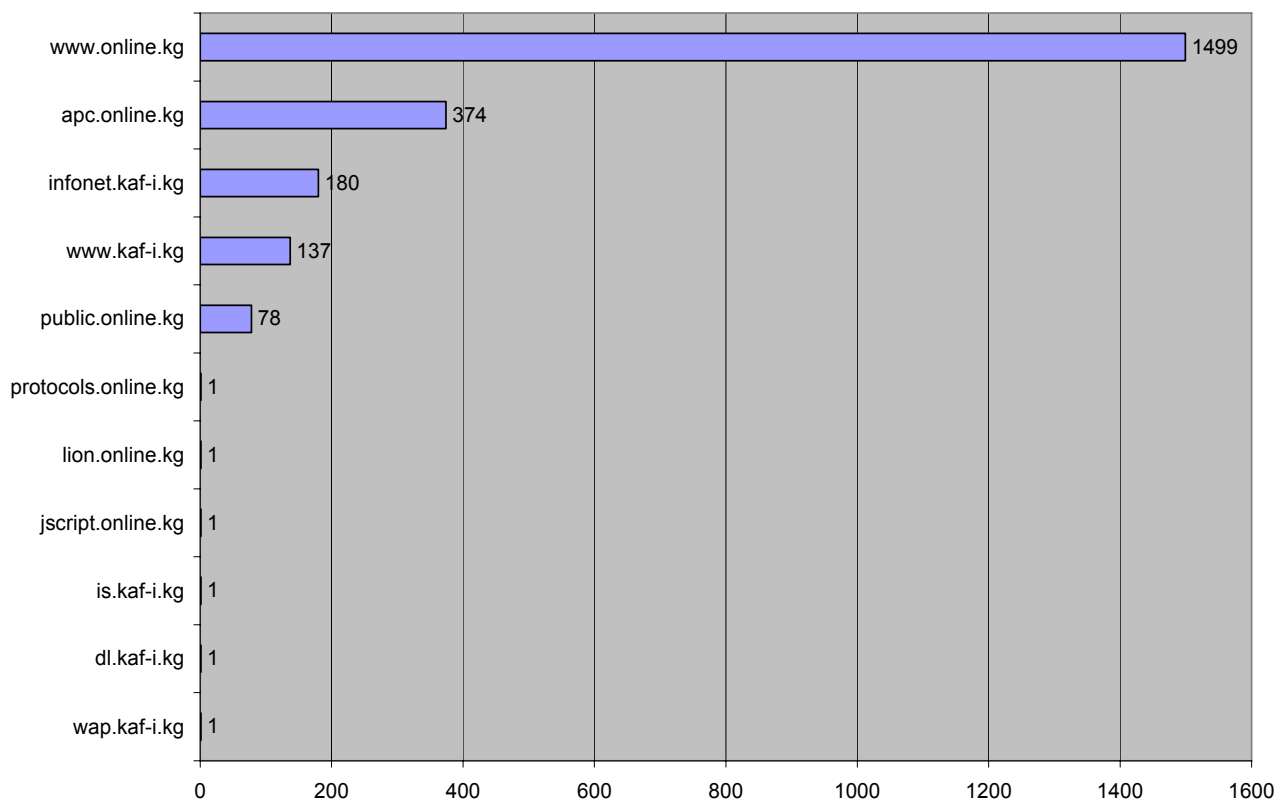


Рис. 13. Распределение документов по веб-сайтам

Сервера с одной проиндексированной страницей не представляют интереса при анализе и не будут рассматриваться по отдельности, поскольку оценки C_p , CD , K_m гипертекста из одного документа равны нулю. Также имеет смысл отдельно рассмотреть характеристики группы из 5 веб-сайтов, имеющих более 1 документа, так как остальные могут оказаться просто не проиндексированными сайтами. Ожидается, что характеристики

сайта www.online.kg будут оказывать доминирующее влияние на характеристики всего множества, так как его документы составляют более половины общего количества.

3.3. Результаты расчетов

Для расчета характеристик была составлена программа на языке T-SQL (см. Приложение I), реализующая расчет матрицы расстояний волновым алгоритмом, преобразование матрицы и расчет описанных в подразделе 2.4 значений (Рис. 14).

Веб-сайт	Длит. Расчетов	Кол-во документов	Кол-во ссылок	K_m	CD	Ср
public.online.kg	3 сек	78	186	52.51%	251 530	0.4630
www.kaf-i.kg	37 сек	137	1543	7.30%	246 444	0.9034
infonet.kaf-i.kg	1 мин 48 сек	180	2811	25.42%	1 530 060	0.7362
arc.online.kg	5 мин 34 сек	374	2041	47.44%	25 008 733	0.5207
www.online.kg	1 ч 31 мин 25 сек	1499	5981	88.11%	2 966 549 445	0.1187
Все крупные	10 ч 30 мин 33 сек	2268	12710	91.71%	10 696 205 152	0.0827
<i>Все множество</i>	11 ч 58 сек	2274	12721	91.73%	10 783 496 902	0.0826

Рис. 14. Сводная таблица результатов расчетов

Прежде всего отметим, что *длительность расчетов* возрастает крайне резко с увеличением количества документов в множестве, поскольку размер матрицы расстояний растет пропорционально квадрату количества узлов графа. Это доказывает, что язык SQL и использование СУБД в качестве инструмента вычислений *нерационально*.

Среди отдельных веб-сайтов *доля отсутствующих путей* больше всего у www.online.kg, что говорит о наличии больших проблем в ссылочном множестве. Наименьшее значение K_m имеет сервер www.kaf-i.kg, что показывает хорошую достижимость среди документов сайта. У остальных серверов значение коэффициента колеблется в широких пределах, однако любое отличное от нуля значение – повод задуматься об улучшении связности системы и внимательно изучить отсутствующие пути.

Как и ожидалось, *преобразованное расстояние* является показательным только при сравнимых масштабах множества документов, в нашем случае это три сайта: public.online.kg, www.kaf-i.kg, infonet.kaf-i.kg. Показательно, что у сайта с меньшим множеством документов (www.kaf-i.kg, 137 документов) преобразованное расстояние CD меньше, чем у public.online.kg (78 документов). Это является подтверждением тезиса (п. 2.4.1) о том, что CD является показателем качества ссылочной структуры гипертекста при сравнимых

масштабах множеств документов. Для сайта infonet.kaf-i.kg наблюдается резкий скачок преобразованного расстояния за счет ссылок длиной 2 перехода, в противоположность www.kaf-i.kg, на котором доля прямых ссылок больше.

Индекс информационной компактности отразил преобразованное расстояние в единой шкале [0;1]. Примечательным является наблюдение, что доля отсутствующих путей и индекс информационной компактности для всех проанализированных множеств оказались в отношении $C_p=1-K_m$ с погрешностью не более 2,5%. Это говорит о том, что в исследованных реальных системах отсутствующие пути оказали подавляющее влияние на оценку гипертекста. Такое сильное влияние отсутствующих путей приводит к необходимости на первых этапах оптимизации ссылочного множества обеспечивать достижимость всех документов, минимизируя K_m , и только после этого переходить к повышению информационной компактности – оптимизации по C_p .

3.4. Рекомендации по применению метода

Апробация метода оценки по нескольким критериям показала, что оптимизация по разным критериям имеет смысл в определенном порядке. Резюмируем выполненную работу, указав последовательность действий, которую предлагается выполнить лицу, ответственному за построение ссылочного множества, чтобы улучшить качество конкретной гипертекстовой системы.

Прежде всего, необходимо обеспечить **автоматизированный сбор** исходной информации об анализируемой системе, поскольку оптимизация происходит итерационно, несколькими этапами, в начале каждого из которых нужно получать обновленную информацию о документах и ссылках в системе. Инструментом сбора такой информации являются так называемые «пауки» (crawlers, краулеры), задача которых – пройти по всем документам и ссылкам гипертекста. Здесь нужно внимательно ограничить множество документов обхода, так это серьезно влияет на характеристики системы (см. п. 2.3.3). Результатом работы краулера должен быть список документов и ссылок между ними. Программа-паук является достаточно простой в разработке и реализовать ее можно достаточно быстро небольшими затратами, даже самостоятельно.

Далее необходимо иметь **инструмент расчета** параметров гипертекста. Как говорилось выше, для этого лучше всего разработать приложение на языке высокого уровня, но, для небольших гипертекстовых множеств, можно воспользоваться разработанной в этой работе программой на языке SQL. На вход инструмента расчетов подаются данные, полученные от краулера, в нашем примере с СУБД необходимо поместить информацию о документах и ссылках в соответствующие таблицы базы данных.

После этого осуществляется первый расчет характеристик системы и их анализ. На первом этапе необходимо внимательно изучить коэффициент K_m и образующие его **отсутствующие пути**, так как при их наличии в гипертексте оценка информационной компактности как отражения длины ссылок, затруднена. Необходимо обеспечить взаимную достижимость всех пар документов в системе, но это не значит создание прямых ссылок для каждой пары документов, необходимо снабдить ссылками лишь некоторые пары и добиться достижимости через промежуточные узлы. Очень популярной является повсеместная ссылка на корневой документ веб-сайта, являющаяся хорошим средством повышения достижимости между документами.

После этого выполняется следующая итерация обхода системы и расчета ее характеристик. Поскольку отсутствующие пути исключены, **индекс компактности** и преобразованное расстояние становятся хорошими показателями длины путей в системе. Если C_p находится на низком уровне (менее 75%), то необходимо изучить распределение длин путей в системе, локализовать чрезмерно длинные пути и обеспечить в гипертексте более короткий путь между соответствующими документами.

Процесс расчета индекса компактности и корректировки на его основании необходимо **повторять** до тех пор, пока коэффициент не войдет в рекомендованные пределы (около 75%). Во время оптимизации по C_p необходимо сохранять $K_m=0$, не упуская это из виду, поскольку отсутствующие пути оказывают радикальное влияние на значение индекса компактности.

В ситуации, когда обеспечить полную достижимость в системе по каким-то соображениям невозможно, рекомендуется принимать решения на основании распределения длин путей, а также уделять больше внимания мостам графа, истокам и стокам (см. подраздел 2.5).

ЗАКЛЮЧЕНИЕ

В работе сформулирована *проблема потери контроля* над структурой разрастающихся гипертекстовых систем и необходимость анализа их ссылочной структуры в целях улучшения пользовательских характеристик системы. В качестве решения предложен метод оптимизации ссылочного множества по нескольким критериям, поскольку оптимальность гипертекста зависит от множества факторов.

В качестве *рекомендации по сбору информации* для анализа предложено использовать утилиты-краулеры, осуществляющие обход системы документов и ссылок и выдающие результат своей работы в удобном к обработке виде, лучше всего таблицы БД. Наличие средства автоматизированного сбора информации о реальной системе значительно ускоряет проведение раундов оптимизации, позволяет увидеть и оценить ее эффект.

В качестве *формальных критериев оптимальности* предложены: длина путей в графе, доля отсутствующих путей, индекс информационной компактности. Метод предлагает использовать критерии не одновременно, а в определенной последовательности и проводить оптимизацию в несколько раундов. Подчеркиваем, что формальные критерии являются лишь средством поддержки принятия решений оптимизации, и не гарантируют реального удобства системы в использовании, ввиду многофакторности и субъективности оценки гипертекста читателем.

Разработанный метод был *опробован на множестве реальных документов*, а также на ряде их подмножеств (отдельных веб-сайтов), пробная программная реализация метода исполнена на языке T-SQL. В результате были вычислены характеристики множеств и выявлены очевидные проблемы в связности документов, позволяющие дать рекомендации по изменениям в системах. Лучшим по связности среди проанализированных сайтов оказался сервер КАФ-Интернет (www.kaf-i.kg), худшим – Kyrgyzstan On-Line (www.online.kg).

Применение данного метода разработчиками гипертекстовых систем позволит сократить вероятность ухода пользователя от чтения гипертекста из-за проблем в структуре ссылок. В системе с оптимизированной навигацией читатель может быстрее находить интересующую его информацию и повысит субъективную оценку системы пользователем. Повышение качества дает преимущества в конкуренции с другими разработчиками на современном, весьма насыщенном, рынке гипертекстовых систем.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

Печатные источники

1. Ицик Бен-Ган, Любор Колар Транзитивное замыкание // WindowsIT Pro. – 2006. – Май. – с. 56-60
2. Похилько А. Создание полнотекстовой поисковой машины для сервера ONLINE.KG: Квалификационная работа бакалавра, 2004. – 50с.
3. Сэдживик Р. Фундаментальные алгоритмы на C++. Алгоритмы на графах. – СПб.: ООО "ДиаСофтЮП", 2002. – 496с.
4. Э. Таненбаум, М. Ван Стеен Распределенные системы. Принципы и парадигмы. – СПб.: Питер, 2003. – 877с.
5. Botafogo, R.A., Rivlin, E., Shneiderman, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. – Maryland: ACM, 1992. – 40с.

Электронные источники

6. А.Чернобаев. *Алгоритмы решения некоторых теоретико-графовых задач*. По материалам лекций МГУ.
(http://www.caravan.ru/~alexch/graphs/graph_alg.htm)
7. Антонов А.С., *Новый подход к межпроцедурному анализу программ*, Автореф. дис. канд. физ-мат. наук, МГУ 1999.
(<http://www.ict.edu.ru/ft/002072/autore.html>)
8. Бочканов Сергей, Быстрицкий Владимир, *Путь с минимальным количеством промежуточных вершин*.
(<http://alglib.sources.ru/graphs/shortpathwave.php>)
9. Г.Е.Кедрова, *Методы оптимизации компьютерной обучающей среды по лингвистике для систем дистанционного обучения в Интернете*.
(<http://www.philol.msu.ru/~kedr/kedr-ulj.htm>)
10. Известия науки, *Интернет-бум возвращается*.
(http://www.kreditpilot.com/html/news/10_06news1.htm)
11. Полковников Е.В. *Показатель суммарной содержательной ценности информационных блоков в задаче оптимизации ссылочных структур гипертекстовых документов*, Конференция молодых ученых по математике, математическому моделированию и информатике.
(http://www.ict.nsc.ru/ws/show_abstract.dhtml?ru+32+2843)

12. *Представление знания в науке, мозгу и компьютере.*
(<http://www.crealab.ru/know.htm>)
13. *Указатель базовых терминов и определений Теории Графов.*
(<http://mportal.narod.ru/Graphs/Search.html>)
14. Эдуард Пройдаков. *История Гипертекста.* Виртуальный компьютерный музей.
(<http://www.computer-museum.ru/histsoft/hypertxt.htm>)
15. Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener. *Graph structure in the web*, WWW9 Conference Proceedings.
(<http://www9.org/w9cdrom/160/160.html>)
16. Barry Schwartz, *Стратегия элементов внутренней ссылочной структуры сайта.*
(http://www.webmasterpro.com.ua/pro/29/1269_1.html)
17. Ian Jacobs, *About the World Wide Web Consortium (W3C).*
(<http://www.w3c.org/Consortium/>)
18. Search Engine Watch Team, *Search Engine Size Wars V Erupts.*
(<http://blog.searchenginewatch.com/blog/041111-084221>)

ПРИЛОЖЕНИЯ

Приложение I. SQL-программа расчета характеристик

```

-- =====
-- ПРОГРАММА РАСЧЕТА ХАРАКТЕРИСТИК ГРАФА ГИПЕРТЕКСТА
-- =====
declare @t datetime
set @t=getdate() -- замер времени
declare @server varchar(100)
set @server='% ' -- анализируемый сервер

-- =====
-- подготовка матрицы
-- =====
if exists (select * from dbo.sysobjects where id =
object_id(N'[dbo].[matrix]') and OBJECTPROPERTY(id,
N'IsUserTable') = 1)
drop table [dbo].[matrix]

CREATE TABLE [dbo].[matrix] (
    [IDfrom] [int] NOT NULL ,
    [IDto] [int] NOT NULL ,
    [length] [int] NULL,
    [path] varchar(7900) NULL
)

-- create matrix
print convert(varchar(100), getdate(),
120)+CHAR(9)+'Подготавливаем матрицу'
insert into matrix (IDfrom, IDto, length)
select P1.ID, P2.ID, -1 from web_pages P1
cross join web_pages P2
inner join web_servers S1
    on P1.serverID=S1.ID
inner join web_servers S2
    on P2.serverID=S2.ID
where S1.server like @server
and S2.server like @server

-- self-loops
print convert(varchar(100), getdate(),
120)+CHAR(9)+'Пути длиной 0 (петли)'
update matrix set length=0, path='' where Idfrom=IDto

-- direct links
print convert(varchar(100), getdate(),
120)+CHAR(9)+'Пути длиной 1'
update matrix set length=1, path=cast(L.IDfrom as
varchar(10))+'+'+cast(L.IDto as varchar(10))
from matrix M
inner join web_links L

```

```

        on M.IDfrom=L.IDfrom and M.IDto=L.IDto
where length = -1

-- index matrix
ALTER TABLE [dbo].[matrix] WITH NOCHECK ADD
    CONSTRAINT [PK_matrix] PRIMARY KEY NONCLUSTERED
    (
        [IDfrom],
        [IDto]
    )

CREATE CLUSTERED INDEX [matrix36] ON [dbo].[matrix]
([length] ASC )
CREATE NONCLUSTERED INDEX [matrix3] ON [dbo].[matrix]
([IDto] ASC )
CREATE STATISTICS [hind_2053582354_2A_3A] ON
[dbo].[matrix] ([idto], [length])
CREATE STATISTICS [hind_2053582354_1A_2A_3A] ON
[dbo].[matrix] ([idfrom], [idto], [length])
CREATE STATISTICS [hind_2053582354_2A_1A_3A] ON
[dbo].[matrix] ([idto], [idfrom], [length])

-- ===== --
-- расчет матрицы расстояний --
-- ===== --
declare @cur int

declare @rc int
set @rc=1
while (@rc>0)
begin
    select @cur=max(length)+1 from matrix where length >
0
    print convert(varchar(100), getdate(),
120)+CHAR(9)+'Пути длиной'+char(9)+cast(@cur as
varchar(100))

    update M
    set M.length=@cur, M.path= cast(M2.IDfrom as
varchar(10))+ '-' +M3.path
    from matrix M -- 1 любая ячейка
    inner join matrix M2
        on (M.IDfrom=M2.idfrom)-- строка
        and M2.length = 1 -- заполненные в ней
        and M2.IDto<>M2.IDfrom -- не диагональ
    inner join matrix M3
        on (M.IDfrom=M3.idfrom OR M.IDto=M3.IDto)--
смежные ребра
        and M3.length > 0 -- заполненные
        AND M3.idfrom=m2.idto
    where 1=1

```

```

        and M.Idto<>M.Idfrom
        and M.length = -1 -- 2 которая еще не заполнена
    set @rc=@@rowcount
    set @cur=@cur+1
end

-- ===== --
-- расчет характеристик --
-- ===== --
-- количество узлов
declare @n bigint
select distinct @n=count(IDfrom) from matrix group by
IDfrom
-- количество ссылок
declare @lnk bigint
select @lnk=count(*) from matrix where length=1

-- преобразование матрицы расстояний
update matrix
set length=@n
where length<0
    and IDfrom<>IDto
print convert(varchar(100), getdate(),
120)+CHAR(9)+'Отсутствующие пути
преобразованы:'+char(9)+cast(@@ROWCOUNT as varchar(10))

-- вычисление CD
declare @CD bigint
select @CD=sum(cast(length as bigint)) from matrix

-- количество отсутствующих путей
declare @Km real
set @Km=1
select @Km=cast(count(*) as real)/cast((@n*(@n-1)) as
real)
from matrix
where length=@n

-- индекс компактности
declare @tmp1 bigint
select @tmp1=(@n*@n)*(@n-1)

declare @tmp2 bigint -- числитель
select @tmp2=@tmp1-@CD

declare @tmp3 bigint -- знаменатель
select @tmp3=@tmp1+@n

declare @CP real
select @CP=cast(@tmp2 as real)/cast(@tmp3 as real)

```

```
-- вывод результатов
print convert(varchar(100), getdate(),
120)+CHAR(9)+'Расчеты окончены,
сек:'+CHAR(9)+cast(datediff(ss,@t, getdate()) as
varchar(10))
print '          Количество узлов:'+char(9)+cast(@n as
varchar(10))
print '          Количество ссылок:'+char(9)+cast(@lnk
as varchar(10))
print '      Доля отсутствующих
путей:'+char(9)+cast((@Km*100) as varchar(10))+ '%'
print ' Преобразованное расстояние:'+char(9)+cast(@CD
as varchar(100))
print '          Индекс компактности:'+char(9)+cast(@CP
as varchar(10))
```