

Кыргызский Национальный Университет им. Ж. Баласагына
Институт Интеграции Международных Образовательных Программ

КЫРГЫЗСКО-АМЕРИКАНСКИЙ ФАКУЛЬТЕТ
КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ И ИНТЕРНЕТ

УДК 681.3 (575.3) (043)

На правах рукописи

Похилько Андрей Федорович

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

Направление: 552800 Информатика и вычислительная техника
Специализация: Компьютерные информационные системы и Интернет

АВТОРЕФЕРАТ
диссертации на соискание квалификационной
академической степени магистра техники и
технологии

Работа выполнена в Институте Интеграции Международных
Образовательных Программ Кыргызского Национального Университета
им. Ж. Баласагына.

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

Рецензент: кандидат технических наук
Асташкевич Б.А.

Ведущая организация: ИИМОП КНУ им. Ж. Баласагына

Защита состоится 29 мая 2006 г. в 14:00 часов на заседании
Государственной Аттестационной Комиссии по защите диссертаций в
Институте Интеграции Международных Образовательных Программ
Кыргызского Национального Университета им. Ж. Баласагына.

С диссертацией можно ознакомиться в Институте Интеграции
Международных Образовательных Программ Кыргызского
Национального Университета им. Ж. Баласагына.

Автореферат разослан «__» мая 2006 г.

Технический секретарь ГАК _____

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Об **актуальности** тематики проведенного 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.

Цель 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 критериев качества ссылочной структуры;
- 4) апробация выработанных критериев на реальных гипертекстовых системах.

Основным **объектом исследования** явились критерии оптимальности внутренней ссылочной структуры системы гипертекстовых документов. Исследовалось использование математической теории графов применительно к формализации и анализу гипертекста.

По тематике исследования существует ряд работ на пространстве СНГ и за рубежом. 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.

Применение предложенного 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.

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.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Диссертация общим объемом 45 страниц состоит из введения, трех основных разделов, заключения, списка источников и приложений.

Во введении обозначена важность темы исследования в связи со второй волной повышения активности веб-разработок, аргументирована необходимость в формализованных методах оценки гипертекстовых систем, поставлены цели и задачи исследования, определена его актуальность и научная новизна.

После 2004 года наблюдается вторая волна “бума” строительства веб-сайтов, проявляющаяся в появлении все новых систем, предоставляющих уже не просто услуги информации вроде электронных библиотек и каталогов, но широкий спектр бизнес услуг, в частности, Интернет-магазины и корпоративные веб-порталы. 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 документов 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? 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. 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 [5]. 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.

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 [5]. 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 [1]. Большинство исследований было проведено в 90-х годах прошлого столетия, однако широкого использования в работе коммерческих фирм их результаты не получили.

В первом разделе приводится информация по теории построения веб-систем и применению математической теории графов для анализа структуры гипертекста.

Формально гипертекст определяется как представление текстовой информации в виде сети, в которой читатели получают свободу перемещаться нелинейным образом. Технологически WWW использует для передачи данных инфраструктуру глобальной компьютерной сети Интернет. Основная суть технологии – архитектура клиент-сервер, с особым серверным и клиентским программным обеспечением [4]. Работа веб-технологии проходит в режиме запрос-ответ. Операционные системы клиента и сервера обслуживают потребности приложений в связи посредством каналов Интернет, которые собственно передают данные от клиента к серверу и обратно.

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. 1.

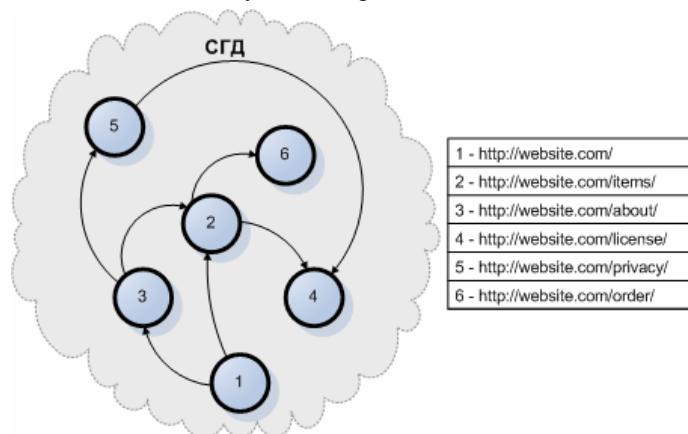


Рис. 1. Схема простой гипертекстовой системы, 6 документов и 7 ссылок

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.

Существенно, что современные программы веб-серверов позволяют строить веб-документы *динамически*. 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.

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.

Второй раздел описывает расчет характеристик формализованной в математический граф гипертекстовой системы, а также ряд критериев оценки оптимальности гипертекста по характеристикам графа.

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.

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. По главной диагонали матрицы смежности всегда стоят нули. На Рис. 2 приведен пример матрицы смежности, соответствующий приведенной выше схеме СГД.

	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. 2. 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 одним из алгоритмов поиска «в ширину», 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. 3).

	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. 3. Матрица расстояний орграфа

Волновой алгоритм является простым и очень надежным итерационным способом вычисления кратчайших путей в графе.

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 может выбираться в зависимости от целей преобразования, с нашей точки зрения будет достаточно брать ее на единицу большей максимально возможной длины пути в графе n . Такой подход дает для недостижимых путей длину большую даже максимально возможной, но все же конечную (Рис. 4).

	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. 4. Преобразованная матрица расстояний

Во втором подразделе второго раздела описан механизм потери контроля над ссылочной структурой гипертекста.

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, однако после перехода к технологическому 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.

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. В этой ситуации необходимо иметь критерии качества ссылочной структуры гипертекстовой системы, расчет которых возможно автоматизировать, а сами они должны быть формальны.

Формальный метод *всесторонней* оценки и сравнения гипертекстовых систем иметь невозможно по причине субъективности, 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*.

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, 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.

Четвертый подраздел второго раздела описывает некоторые критерии оценки ссылочной структуры гипертекста.

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.

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

$$CD = \sum_i \sum_j C_{ij} \quad (1)$$

Наименьшее преобразованное расстояние имеет полный орграф:

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

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

$$CD_{\max} = n \cdot (n^2 - n) \quad (3)$$

Недостатком этого критерия является то, что сравнить два разных гипертекста с его помощью можно, только если в них одинаковое количество документов.

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.

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*:

$$K_m = \frac{Q_m}{n^2 - n} \quad (4)$$

Значение доли отсутствующих путей колеблется в пределах [0;1] и допускает сравнение систем гипертекстовых документов между собой

В качестве еще одного критерия оптимальности гипертекстовой системы предлагается использовать индекс компактности, отражающий пересеченность гипертекстовой системы ссылками и близость графа системы к полному. Индекс компактности описан в работе Ботафого [5] и выражается как

$$C_p = \frac{CD_{\max} - CD}{CD_{\max} - CD_{\min}} \quad (5)$$

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

$$C_p = \frac{n^3 - n^2 - \sum_i \sum_j C_{ij}}{n^3 - 2n^2 + n} \quad (6)$$

Индекс компактности опирается на понятие *преобразованного расстояния* CD , но в отличие от него изменяется в пределах [0;1] для любого графа, давая, таким образом, возможность сравнивать различные гипертексты между собой.

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.

Третий раздел описывает эксперимент по применению критериев для оценки реальных систем.

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 и была получена в формате СУБД MySQL версии 4.2, так как поисковая машина КАФ-Интернет используется 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

Прежде чем приступить к расчетам была осуществлена проверка, что исходные данные не содержат противоречий и ненужного «мусора». В итоге очистки и подготовки данных получена БД в Microsoft SQL Server, описывающая 11 веб-серверов, 2274 страниц, 12721 ссылки. Сервера с одной проиндексированной страницей не представляют интереса при анализе и не рассматривались по отдельности, поскольку оценки C_p , CD , K_m гипертекста из одного документа равны нулю.

Для эксперимента была составлена программа на языке T-SQL, реализующая заполнение матрицы расстояний волновым алгоритмом, преобразование матрицы и расчет значений коэффициентов (Рис. 5).

Веб-сайт	Длит. Расчетов	Кол-во документов	Кол-во ссылок	Km	CD	Cr
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
apc.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
Все множество	11 ч 58 сек	2274	12721	91.73%	10 783 496 902	0.0826

Рис. 5. 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.

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 $Cr=1-K_m$ 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 Cr .

Апробация метода оценки показала, что оптимизация по разным критериям имеет смысл в определенном порядке. 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. Далее

необходимо иметь *инструмент расчета* параметров гипертекста, 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.

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 путей, 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 Cr 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 (более 75%). Во время оптимизации по Cr необходимо сохранять $K_m=0$, не упуская это из виду, поскольку отсутствующие пути оказывают радикальное влияние на значение индекса компактности.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

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 выдающие результат своей работы в годном к обработке виде, лучше всего в виде таблиц 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 – Кыргызстан 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 ЛИТЕРАТУРЫ

1. 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>)
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. 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>)
6. Botafogo, R.A., Rivlin, E., Shneiderman, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. – Maryland: ACM, 1992. – 40с.

РЕЗЮМЕ

ПОХИЛЬКО Андрей Федорович

Анализ оптимальности информационных структур в системах гипертекстовых документов

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

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.

Цель работы – разработка метода анализа оптимальности ссылочных структур в гипертексте и их оптимизации на основе формальных критериев, объектом исследования являются критерии оптимальности ссылочной структуры гипертекста.

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.

РЕЗЮМЕ

ПОХИЛЬКО Андрей Федорович

Гипертексттик документтердин системасында маалымат структураларынын оптималдуулугун анализдөө

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

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.

Изилдөөнүн объекти: Гипертексттин таянуу структурасынын оптималдуулугун өлчөөдөгү критерийлер

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

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.

RESUME**Andrey F. POHILKO****The analysis of the information structures' optimality in the system of the hypertext documents****Keywords:** hypertext, structure, graph theory, criteria, optimization

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

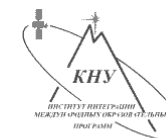
Purpose of the work – the development of the analysis method of the hyperlink structures optimization in the hypertext and theirs optimization based on the formal criteria.

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



ИИМОП КНУ, Похилько А.Ф.

Анализ оптимальности информационных структур в системах гипертекстовых документов
Бишкек 2006