

Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**«Санкт-Петербургский государственный университет
промышленных технологий и дизайна»
Высшая школа технологии и энергетики**

П. Е. Антонюк

ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ

Учебно-методическое пособие

Утверждено Редакционно-издательским советом ВШТЭ СПбГУПТД

Санкт-Петербург
2022

УДК 681.3(075)

ББК 73я7

А 724

Рецензенты:

заведующий кафедрой ПМИ ВШТЭ СПбГУПТД, кандидат технических наук

В. П. Яковлев;

доцент кафедры информационно-измерительных систем и технологий

СПбГЭТУ «ЛЭТИ», кандидат технических наук

Н. В. Орлова

Антонюк, П. Е.

А 724 Информационно-поисковые системы: учебно-методическое пособие /
П. Е. Антонюк. – СПб.: ВШТЭ СПбГУПТД, 2022. – 45 с.

Учебно-методическое пособие соответствует программам и учебным планам дисциплины «Информационно-поисковые системы» для студентов, обучающихся по направлению подготовки 01.03.02 «Прикладная математика и информатика». Учебно-методическое пособие носит обзорный характер и содержит лекции и лабораторные работы.

Пособие предназначено для подготовки бакалавров очной формы обучения.

© Антонюк П. Е., 2022

© ВШТЭ СПбГУПТД, 2022

СОДЕРЖАНИЕ

Содержание и тематика дисциплины	4
ТЕМА 1. ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ.....	5
Лабораторная работа 1. Исследование эффективности поиска.....	7
ТЕМА 2. ПОИСКОВЫЕ СИСТЕМЫ И КАТАЛОГИ	10
Лабораторная работа 2. Составление и отладка тематического запроса	17
ТЕМА 3. СПЕЦИАЛЬНЫЕ ВОПРОСЫ ИНФОРМАЦИОННОГО ПОИСКА	18
Лабораторная работа 3. Использование пользовательского поиска в Internet	22
ТЕМА 4. АНАЛИЗ УДОБСТВА (USABILITY) ПОИСКА ДЛЯ ИНФОРМАЦИОННОГО САЙТА	24
Лабораторная работа 4. Оценка качества Интернет-ресурса.....	30
ТЕМА 5. АНАЛИЗ СПРОСА НА ТОВАРЫ САЙТА СРЕДИ АУДИТОРИИ ПОИСКОВЫХ СИСТЕМ.....	32
Лабораторная работа 5. Составление запросов на поиск информации в сети Internet	34
ТЕМА 6. КЛАССИФИКАЦИОННЫЕ СИСТЕМЫ ПОИСКА ДОКУМЕНТОВ	37
Лабораторная работа 6. Рынок электронных сделок, представленный системой заказа товара и услуг	43
Библиографический список	45

СОДЕРЖАНИЕ И ТЕМАТИКА ДИСЦИПЛИНЫ

Решение задач современного постиндустриального общества невозможно без использования информационных технологий, применяющихся во всех сферах человеческой деятельности. Информационные технологии немыслимы без информации, которую необходимо найти и структурировать.

Информационный поиск отражает стремление человека к нахождению необходимых знаний. Структуры систем, предназначенных для поиска информации, совершенствуются с каждым поколением вычислительной техники и программных средств.

Цель изучения дисциплины «Информационно-поисковые системы» – приобретение студентами знаний в области методов проектирования и эксплуатации информационно-поисковых систем на основе современных систем управления базами данных.

Задачи дисциплины – научить студентов:

- работать с многофункциональными программными комплексами для управления предприятием;
- использовать информационно-поисковые системы для поиска информации;
- принципам применения современных информационно-поисковых систем в профессиональной деятельности.

Настоящее учебно-методическое пособие содержит лекции, позволяющие изучить историю информационно-поисковых систем, их классификацию и примеры использования в современных информационных средах и устройствах. После каждой лекции приведен текст лабораторной работы, выполняющейся по итогам прохождения соответствующей темы.

ТЕМА 1. ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Информация – любые сведения, являющиеся объектом хранения, обработки и передачи. Любая информация может быть структурирована или не структурирована.

«Потребность в системах, ориентированных на накопление и эффективную обработку неструктурированной или слабоструктурированной информации, привела к возникновению в 70-х годах XX века отдельной ветви программного обеспечения – систем управления базами данных (СУБД), на основе которых создавались первые информационные системы. Однако развитые коммерчески успешные информационные системы, ориентированные на накопление и обработку текстовых документов, получили распространение только в конце 1980-х – начале 1990-х годов прошлого века. Дадим определение подобных систем. Информационно-поисковая система (ИПС, англ. Information retrieval system – IRS) – информационная система с единым хранилищем единичных элементов данных и с развитым инструментарием их поиска и последующего отбора.

По виду единичных элементов данных ИПС делятся на *фактографические* и *документальные*. *Фактографические ИПС* – информационная система, в которых единичным элементом данных, имеющим отдельные смысловые значения, является *запись*, образуемая конечной совокупностью полей-атрибутов. Так, например, в СУБД MSAccess запись может состоять из следующих полей:

Таблица 1.1 – Пример записи в MS Access

Табельный номер	Фамилия	Имя	Отчество	Год рождения	Оклад, руб.
971879	Иванов	Сергей	Петрович	1970	45000

Работа с фактографическими ИПС требует либо изначально структурированных данных, либо предварительной структуризации данных. Из этого вытекают *недостатки* фактографических ИПС:

- 1) структуризация данных требует больших накладных расходов, приводящих к материальным издержкам;
- 2) извлечение из текста данных по формализованным позициям для ввода в фактографические ИПС может приводить к ошибкам и потере части информации, которая в исходных источниках имеется, но из-за отсутствия в базе данных соответствующих элементов в них не может быть отражена.

Документальные ИПС – информационные системы, единичным элементом которых является неструктурированный на более мелкие элементы документ, такой как текстовый файл, звуковой или графический файл.

Основной задачей документальных ИПС является накопление и представление пользователю документов, которые по содержанию, тематике, реквизитам и прочее соответствуют его информационным потребностям.

Информационная потребность – осознанное понимание различия индивидуального знания, определяемое разностью между субъективным восприятием предмета деятельности и уровнем знаний об этом предмете, накопленных обществом.

Важным понятием, связанным с информационной потребностью, является *пертинентность* – соответствие найденных документов информационным потребностям пользователя.

В зависимости от особенностей реализации хранилища документов и механизмов поиска документальные ИПС делятся на *системы на основе индексирования* и *семантически-навигационные системы*. В первых из них исходные документы помещаются в базу без дополнительного преобразования, но смысловое содержание каждого документа отображается в некотором *поисковом пространстве*. Процесс подобного отображения заключается в присвоении каждому документу некоторого *индекса* – координаты в поисковом пространстве. Это процесс называется индексированием, его ещё можно описать как описание содержания документа с помощью формализованного информационного языка. В семантически-навигационных системах документы, помещаемые в базу документов, оснащаются специальными навигационными конструкциями, соответствующими смысловым связям между различными документами или отдельными фрагментами одного документа. Поиск осуществляется путем явной навигации по смысловым связям между документами. Такой подход реализован в настоящее время в гипертекстовых ИПС.

Поисковым образом документа (ПОД) называется формализованное описание индекса документа.

Поисковым образом запроса (ПОЗ) к базе документов называется выражение пользователем своих информационных потребностей средствами и языком поискового пространства.

Система на основе определенных критериев и способов ищет документы, ПОД которых соответствуют или близки ПОЗ, и выдает соответствующие документы.

К показателям эффективности функционирования информационно-поисковой системы относятся:

1. *Полнота информационного поиска* R – отношение найденных пертинентных документов A к общему числу пертинентных документов C , имеющих в системе или исследуемой совокупности документов $R=A/C$.

2. *Точность информационного поиска* P – отношение числа найденных пертинентных документов A к общему числу документов L , выданных на запрос пользователя $P=A/L$.

3. *Информационный шум системы* – наличие среди отобранных на запрос пользователя нерелевантных документов. Коэффициент информационного шума K определяется отношением числа нерелевантных документов $(L-A)$, выданных в ответе пользователю, к общему числу документов L , выданных на запрос пользователя $K=(L-A)/L$.

В идеале показатели 1) и 2) должны приближаться к единице, хотя на практике их значения колеблются в пределах 0,6 – 0,9.

Лабораторная работа 1. Исследование эффективности поиска

Цель работы

Получить основные навыки поиска информации в Интернете, изучить популярные поисковые системы, определить (сформулировать) критерии эффективности поиска в них и определить наиболее эффективную поисковую систему в Интернет.

Порядок выполнения работы

Изучить возможности следующих поисковых систем Интернет:

<http://www.google.com>

<http://www.rambler.ru>

<http://www.mail.ru>

<http://www.yandex.ru>

Примечание: список исследуемых поисковых систем может быть расширен.

При изучении функций и возможностей исследуемых поисковых систем следует использовать доступную документацию и справочные системы.

Принять список критериев оценки эффективности поисковых систем.

Предлагаются следующие критерии:

Полнота поиска.

Точность поиска.

Назначить весовые коэффициенты (веса) принятым критериям. Весовой коэффициент – параметр, отражающий значимость, относительную важность, «вес» данного критерия в сравнении с другими критериями. Например, для точности поиска весовой коэффициент принимаем равным 0,8, для полноты поиска – 0,2 (т.е. точность «весомее» или важнее, чем полнота, потому что 0,8 больше, чем 0,2). Сумма всех весов должна быть равной 1 ($0,8 + 0,2 = 1$, а может быть $0,6 + 0,4 = 1$; если критерия выбрано три, то например, $0,3 + 0,2 + 0,5 = 1$).

Принять список тем запросов для оценки эффективности поисковых систем.

Предлагаются следующие темы:

Информационные системы.

Вычислительная техника.

Информационные технологии.

Бизнес-информатика.

Кибернетика.

Электроника.

Прикладная математика и информатика.

Образование.

Менеджмент и управление.

Экспертные системы.

Информационная безопасность.

Искусственный интеллект.

Мультимедийные технологии.

Нано-технологии.

Телекоммуникации.

Сформулировать запросы по каждой теме. Запрос должен включать в себя несколько (не менее двух) ключевых слов и относиться к данной теме. Примеры запросов: «сетевые технологии», «защита информации», «виртуальная реальность в социальной сфере» и т.п.

Выполнить каждый запрос в каждой исследуемой поисковой системе.

Если результаты поиска в какой-либо поисковой системе Вас не удовлетворили, сформируйте более точный запрос (изменив или добавив ключевые слова), позволяющий найти исчерпывающую информацию об объекте. Повторите поиск в каждой поисковой системе.

Из полученных списков результатов выбрать следующую информацию:

Общее количество найденных документов (Д).

Количество релевантных документов различной ценности (РД).

При определении количества релевантных документов должно быть использовано понятие «релевантность» или «пертинентность». Количество релевантных документов оценивается при просмотре текста первых 10 найденных документов. При этом определяется ценность найденной информации (степень удовлетворения найденным документом Ваших информационных потребностей). Ценность информации определяется по 3-балльной шкале: 2 – имеет ценность, 1 – имеет частичную ценность, 0 – не имеет ценности.

Результаты выполнения запросов свести в таблицу 1.2

Таблица 1.2 – Результаты выполнения запросов

Тема	Google		Mail.ru		Rambler		Yandex	
	Д	РД	Д	РД	Д	РД	Д	РД
	2		1		0		2	
Информационные системы								
Вычислительная техника								
...								

Выполнить первичную обработку результатов, сведя их в таблицу 1.3

Вычислить средние арифметические значения показателей Д, РД(2), РД(1) и РД(0) для каждой поисковой системы.

Определить место каждой поисковой системы по критерию «Полнота поиска». Для определения места использовать среднее количество найденных документов Д. Система, которая нашла больше документов, считается лучшей. Самой лучшей системе присваивается место 1, самой худшей – место N (N – количество исследуемых систем).

Для каждой поисковой системы ЭБ определить коэффициент точности поиска Р по следующей формуле:

$$P = \frac{a}{a+b},$$

где a – число релевантных документов, выданных поисковой системой в ответ на запрос;

$$a = \text{РД}(2) + 0.5 * \text{РД}(1);$$

b – число нерелевантных документов, выданных поисковой системой в ответ на запрос, $b = \text{РД}(0)$.

Определить место каждой поисковой системы по критерию «Точность поиска». Система, имеющая большее значение коэффициента точности поиска P , считается лучшей. Самой лучшей системе присваивается место 1, самой худшей – место N (N – количество исследуемых систем).

Вычислить коэффициент поискового шума S по следующей формуле:

$$S = 1 - P.$$

Вычислить сумму мест каждой исследуемой системы R по следующей формуле:

$$R = \sum_{i=1}^m w_i (N - q_i + 1),$$

где m – количество критериев оценки;

i – номер критерия оценки поисковой системы;

w_i – весовой коэффициент для критерия оценки i ;

q_i – место поисковой системы по критерию оценки i ;

N – количество исследуемых систем.

Определить лучшую поисковую систему Интернет из исследуемых.

Оформить результаты работы по предлагаемой форме.

Оформление результатов

В качестве предварительной информации привести заполненную таблицу 1.2.

Основные результаты работы оформить в виде таблицы 1.3.

Таблица 1.3 – Результаты сравнительного анализа поисковых систем Интернет

Критерий	Google	Mail.ru	Rambler	Yandex
Полнота поиска (вес – ...) – среднее количество найденных документов (Д)				
Место (полнота поиска)				
Среднее количество пертинентных документов (РД2)				
Среднее количество частично пертинентных документов (РД1)				

Среднее количество непертинентных документов (РД0)				
Коэффициент точности поиска Р (вес – ...)				
Место (точность поиска)				
Коэффициент поискового шума (S)				
Сумма мест (R)				

ТЕМА 2. ПОИСКОВЫЕ СИСТЕМЫ И КАТАЛОГИ

Объем хранящейся в Интернете информации чрезвычайно велик и оценивается десятками терабайт. По состоянию на 26 сентября 2021 года на серверах сети хранится 1 892 465 450 Web-страниц*. Поиск информации в сети можно вести с помощью поисковых систем (ПС) и каталогов. Мощные поисковые системы и каталоги являются сложными техническими комплексами, содержащими десятки быстродействующих компьютеров, обслуживание которых ведут сотни специалистов.

Вначале рассмотрим принцип действия ПС, а затем – каталогов.

Принцип действия ПС

ПС иначе называют: поисковыми средствами, поисковыми машинами, автоматическими индексами. Эквивалентными иностранными терминами являются: английский – Search Engines, немецкий – Suchmaschinen, французский – Le systeme de prospection.

Работа ПС основывается на формировании запроса, по которому происходит отбор нужных документов из распределенной базы данных, хранящейся на серверах в Интернете. Запрос формируется с помощью ключевых слов (одного или нескольких). Результаты поиска выдаются пользователю в виде списка адресов (гиперссылок) и краткой аннотации к ним.

Ключевое слово – это лексическая единица, являющаяся существительным, прилагательным, глаголом, числительным, наречием или местоимением естественного языка, которая в наибольшей степени отражает содержание всего искомого документа. При формировании запроса могут использоваться не только отдельные ключевые слова, но и словосочетания, состоящие из нескольких ключевых слов.

Запрос – это набор соединенных операторами ключевых слов, с помощью которых поисковая система автоматически ведет поиск и отбор необходимых

*internetlvestats.com

документов. Другими словами, запрос – это инструкция (команда) для ПС на поиск нужных документов.

Запросы бывают двух типов:

- простые запросы состоят из отдельных ключевых слов или словосочетаний;

- сложные (или расширенные, *advanced*): запросы, кроме ключевых слов, содержат логические и другие операторы.

Поиск – это процедура отбора нужных документов, хранящихся в сети.

Поиск осуществляется:

- автоматически с помощью ключевых слов, вводимых в ПС;

- путем последовательного ручного прохода по рубрикам в каталогах.

Поиск с помощью ключевых слов сводится к их вводу (формированию запроса) в специальное поле Поиск и последующему нажатию кнопки Найти. Кнопка НАЙТИ в разных ПС может называться Search, Find, Go, Go Get it, Suchen.

Для ускорения процедуры поиска ПС производит обработку хранящихся в сети документов – индексацию.

Индексация – это процедура автоматического создания базы данных, в которой хранятся ключевые слова, аннотации документов и доменные адреса, по которым размещены эти документы. В базе данных каждому документу ставится в соответствие свой набор ключевых слов и доменных адресов.

Когда идет обслуживание конкретного запроса на поиск нужной информации, то ПС сравнивает ключевые слова, введенные пользователем, с ключевыми словами, полученными в процессе индексации и хранящимися в базе данных ПС. При совпадении этих слов пользователю выдается доменный адрес данного документа, т. е. указывается место его хранения в сети.

Процедура поиска по ключевым словам очень напоминает работу с обычной книгой, в которой есть алфавитный указатель. Читатель с помощью алфавитного указателя определяет номера страниц книги, на которых присутствует нужное ключевое слово. Этим ускоряется поиск нужной информации. В данном случае номер страницы книги – это как бы доменный адрес Web-страницы.

Индексация документов, размещенных на различных серверах, производится поисковыми системами автоматически с помощью специальных программ – роботов (Robots). Другие названия этих программ – пауки (Spiders) или черви (Worms).

Обычно системы работают поэтапно. Сначала поисковый робот получает контент, затем индексатор генерирует доступный для поиска индекс, и наконец, поисковик обеспечивает функциональность для поиска индексируемых данных. Чтобы обновить поисковую систему, этот цикл индексации выполняется повторно.

Поисковые системы работают, храня информацию о многих веб-страницах, которые они получают из HTML-страниц. Поисковый робот или «краулер» (англ. Crawler) – программа, которая автоматически проходит по всем ссылкам, найденным на странице, и выделяет их. Краулер, основываясь на

ссылках или исходя из заранее заданного списка адресов, осуществляет поиск новых документов, ещё не известных поисковой системе. Владелец сайта может исключить определённые страницы при помощи файла robots.txt, используя который можно запретить индексацию файлов, страниц или каталогов сайта.

Поисковая система анализирует содержание каждой страницы для дальнейшего индексирования. Слова могут быть извлечены из заголовков, текста страницы или специальных полей – метатегов. Индексатор – это модуль, который анализирует страницу, предварительно разбив её на части, применяя собственные лексические и морфологические алгоритмы. Все элементы веб-страницы вычлняются и анализируются отдельно. Данные о веб-страницах хранятся в индексной базе данных для использования в последующих запросах. Индекс позволяет быстро находить информацию по запросу пользователя.

Ряд поисковых систем, подобных Google, хранят исходную страницу целиком или её часть, так называемый кэш, а также различную информацию о веб-странице. Другие системы хранят каждое слово каждой найденной страницы. Использование КЭШа помогает ускорить извлечение информации с уже посещённых страниц. Кэшированные страницы всегда содержат тот текст, который пользователь задал в поисковом запросе. Это может быть полезно в том случае, когда веб-страница обновилась, то есть уже не содержит текст запроса пользователя, а страница в кэше ещё старая. Эта ситуация связана с потерей (вымиранием) ссылок (англ. linkrot) и дружественным по отношению к пользователю (юзабилити, англ. usability, см. тему 4) подходом Google. Это предполагает выдачу из кэша коротких фрагментов текста, содержащих текст запроса. Действует принцип наименьшего удивления, пользователь обычно ожидает увидеть искомые слова в текстах полученных страниц (User expectations). Кроме того, что использование кэшированных страниц ускоряет поиск, страницы в кэше могут содержать такую информацию, которая уже нигде более не доступна.

Назначение роботов очевидно: эти программы постоянно «осматривают», сканируют сеть, запоминают, что где лежит, чтобы в нужный момент времени показать пользователю точное место хранения документа (т. е. его доменный адрес). Этим объясняется высокая скорость поиска нужной информации. Фактически роботы начинают поиск задолго до обслуживания конкретного запроса.

Тенденция развития ПС такова, что наиболее быстродействующие ПС стремятся произвести индексацию всего документа, а не только его названия и первых предложений текста. Наиболее совершенные роботы при индексации сканируют не только главную (домашнюю) страницу, но и по гиперссылкам заходят вглубь сайта (узла). Ограничивается такая идеалистическая картина индексации лишь низкой скоростью работы современных ПС и общим быстродействием всей сети.

Предварительная индексация хранящихся в сети документов позволяет впоследствии за несколько секунд обслуживания запроса обработать (отсортировать) гигабайты разнообразной информации.

С помощью механизма предварительной индексации можно получить хороший результат поиска в случаях, когда удастся точно сформировать запрос с помощью небольшого числа ключевых слов. В иных случаях пользователь либо получит большое число ссылок (иногда говорят, линков, т. е. связей), которые являются лишь «информационным шумом», либо вовсе не удастся найти необходимый документ.

Значительно повышают избирательность поиска фильтры, которыми снабжаются поисковые системы.

Фильтры позволяют:

- ограничить список отбираемых документов с помощью логических операторов (выполнить так называемый сложный поиск);
- ограничить отбираемый материал временным отрезком, определенными датами создания разыскиваемого документа (например, между 31 июля 2016 г. и 19 апреля 2017 г.);
- отобрать документы, составленные только на определенном языке (русский, английский);
- ограничить отбор документов территорией размещения серверов (например, только Европа);
- ограничить поиск определенной частью документа (заголовок, доменный адрес);
- отобрать документы, которые содержат фразу с заданным порядком расположения ключевых слов.

Поиск русских документов в сети дополнительно осложняется особенностями национальной грамматики. В русском языке (в отличие от английского языка) необходимо учитывать падежные окончания ключевых слов.

В 1996 году был реализован поиск с учётом русской морфологии на поисковой машине Altavista и запущены оригинальные российские поисковые машины Рамблер (произносится Ремблер, а переводится как бродяга, праздношатающийся, его адрес www.rambler.ru) и Апорт (www.aport.ru). 23 сентября 1997 года была открыта поисковая машина Яндекс (www.yandex.ru).

Русскоязычные ПС позволяют отбрасывать окончания ключевых слов и заменять их метасимволами «*» и «?». При этом на запрос «самар*» будут найдены документы, содержащие слова: Самара, Самары, самарский, самарская и т. п.

Поисковая система Aport использует принципы искусственного интеллекта. Она обрабатывает запрос таким образом, что подбирает синонимы введенным ключевым словам и автоматически решает проблему падежных окончаний.

При поиске информации ПС может делать две ошибки: пропускать (не отбирать) нужные пользователю документы и, наоборот, отбирать (присылать) посторонние документы (мусор, информационный шум). Эти ошибки,

совершаемые автоматическими ПС, описываются терминами «избирательность» и «чувствительность».

Заметим, что в приведенной фразе есть некоторая терминологическая неточность: ПС отбирают не документы, а лишь гиперссылки на них. С помощью полученных ссылок пользователь просматривает или загружает на собственный компьютер, понравившийся ему документ. Однако методически удобнее говорить о несоответствии запросу документа, а не гиперссылки.

Избирательность – способность ПС отбирать документы, соответствующие запросу, не включая лишних документов. Качество избирательности характеризуется числом ошибок первого рода – число отобранных документов, не соответствующих запросу. Чем выше избирательность, тем меньше посторонних документов попадает к пользователю.

Избирательность можно изменять (регулировать) с помощью логических операторов (фильтров).

Чувствительность – способность ПС отбирать документы, соответствующие запросу, не пропуская нужных документов. Чувствительность характеризуется числом ошибок второго рода – числом пропусков нужных документов. Чем выше чувствительность, тем меньше вероятность пропуска нужного документа.

Релевантность – степень (мера) соответствия (адекватности) найденного в процессе поиска документа сделанному запросу.

Уточнить запрос (говорят: отфильтровать информацию) позволяют логические операторы OR, AND, NOT.

Использование логического оператора AND (И) приводит к отбору документов, которые обязательно содержат все перечисленные в запросе ключевые слова, соединенные этим оператором.

Оператор NOT (НЕТ) позволяет исключить документы, которые содержат ключевое слово, указанное после этого оператора.

С помощью оператора NEAR (вблизи, рядом) пользователь может отбирать документы, в которых ключевые слова, соединенные этим оператором, будут находиться поблизости друг от друга, а не в разных концах документа.

Оператор FOLLOWED BY позволяет отбирать документы, в которых ключевые слова следуют друг за другом в заданном порядке.

Оператор ADJ отбирает документы, в которых ключевые слова являются смежными (следуют одно за другим).

К сожалению, каждая ПС имеет собственный синтаксис запросов. Поэтому перед формированием расширенных запросов необходимо уточнить правила использования логических и других операторов.

Современные ПС становятся интеллектуальными. Используя принципы искусственного интеллекта, они ранжируют (располагают) выводимый список документов (ссылок на них) в зависимости от степени их релевантности. При этом ПС анализируют положение найденных ключевых слов в документе (заголовок или текст), число повторений ключевых слов, их взаимное

расположение в документе. Наиболее точно отобранные документы располагаются в начале списка найденных в процессе поиска документов.

Для ранжирования найденных документов используют следующие показатели.

- *Положение ключевого слова на странице* (keyword prominence) – показатель, определяющий, как близко к началу документа находится заданное ключевое слово. Как правило, чем ближе к началу страницы располагается ключевое слово, тем точнее документ соответствует запросу.

- *Частота ключевого слова* (keyword frequency) – показатель, учитывающий абсолютную частоту использования ключевых слов (т. е. сколько раз встречается данное ключевое слово на странице).

- Наибольший «вес» при ранжировании документов имеют слова, расположенные в заголовке Web-страницы (так называемый *титул*). Именно этот заголовок отображается в строке заголовка браузера при просмотре страницы, и приведенные там слова имеют наибольшую ценность для ПС.

- Некоторые поисковые системы при ранжировании учитывают *индекс цитирования* (link popularity) – количество сайтов, которые ссылаются на данный сайт.

Не всякий запрос можно четко сформулировать с помощью небольшого числа ключевых слов. Поэтому, кроме индексного механизма, используются и другие механизмы поиска и хранения информации в сети.

Принцип действия каталогов

Как отмечалось ранее, поиск информации в Интернете можно производить с помощью поисковых систем и каталогов. Другое нередко используемое название поисковых средств, построенных с помощью каталогов, – иерархические ПС.

Термин «иерархия» означает: расположение частей или элементов целого в порядке от высшего к низшему. В данном случае правильнее говорить: иерархия – это уточнение содержания документа от его общих характеристик к частным характеристикам.

Синонимами термина «каталог» являются слова: директория, справочник, категория, рубрикатор.

Каталоги представляют собой тематически подобранные сетевые адреса, которые сопровождаются краткими комментариями (аннотациями).

Каталоги появились самым естественным путем: пользователи отбирали для себя ссылки на любимые места в Интернете и составляли упорядоченные тематические списки с детализацией каждого раздела (рубрики). Именно так возник всемирно известный каталог Yahoo!, созданный студентами Стэндфордского университета Дэвидом Фило и Джерри Янгом.

В каталогах размещение информации ведется с помощью многоуровневой рубрикации, причем на каждом уровне происходит все большая детализация (уточнение) сведений об искомом (хранящемся)

документе. При этом каждая последующая рубрика не исключает свойств документа, определенных предыдущей рубрикой, а лишь уточняет их. Иначе говорят: рубрики находятся в отношении соподчинения, иерархии. Еще говорят: используется принцип последовательного уточнения.

Важным термином, характеризующим работу каталогов, является рубрикация.

Рубрикация – помещение документа в соответствующий раздел (подраздел, рубрику), которое, как правило, производится вручную специалистами (модераторами, аналитиками) или авторами разработанных документов. Разработчиков Web-страниц иногда называют владельцами ресурсов или Web-дизайнерами.

Процедура рубрикации достаточно субъективна и осуществляется на основании индивидуальных представлений людей о данной предметной области. Авторы при размещении своих страниц порой руководствуются не соображением точного соответствия документа названию рубрики, а другими соображениями. Например, с целью увеличения посещаемости своих страниц их помещают в рубрики, наиболее популярные в данный момент времени. Документ, содержащий разнообразную информацию на Web-странице, порой помещают в несколько рубрик.

Перечислим доменные адреса некоторых отечественных каталогов:

<http://www.mail.ru/>, <http://www.km.ru/url/>,

<http://www.ru/>, <http://www.ulitka.ru/>

<http://www.susanin.net/>, <http://students.informika.ru/>,

<http://www.stars.ru/>, <http://yp.piter.com/>.

В Интернете, кроме иерархических каталогов, встречаются и другие типы каталогов – неиерархические, в которых нет соподчинения, например каталоги, использующие алфавитно-предметную классификацию. В качестве примера можно назвать энциклопедии, музыкальные сайты, телефонные справочники и др.

Таким образом, поисковые системы и тематические каталоги имеют свои достоинства и недостатки, взаимно дополняют друг друга.

Поисковые системы «осматривают» большое число сайтов, однако автоматический отбор искомых документов сопровождается достаточно большим числом ошибок.

База данных каталога хранится на одном сервере и содержит меньший объем информации, чем ПС. Тем не менее, результаты поиска с помощью каталогов имеют большую релевантность по сравнению с поиском с помощью ПС.

Результаты поиска с помощью различных ПС будут отличаться друг от друга. Это объясняется тем, что в каждой ПС проиндексировано различное число документов и эти документы совсем не обязательно одинаковые. Кроме того, алгоритмы индексации на каждой ПС имеют свои особенности.

Существуют национальные ПС, которые в первую очередь обрабатывают информацию, составленную на родном языке.

Периодичность обновления информации с помощью роботов также различна, и она определяется быстродействием конкретной ПС (частотой посещения сайтов). Для расширения просматриваемого в сети пространства рекомендуют использовать несколько ПС.

Метапоисковая система позволяет вести автоматический поиск по сделанному запросу с использованием сразу нескольких поисковых машин.

Укажем адреса некоторых метапоисковых систем общего назначения:

www.metacrawler.com, www.motor.com, www.infozoid.com, www.accufind.com.

Кроме ПС общего назначения существуют специализированные ПС.

Они предназначены для поиска

- музыкальных файлов (www.midi.ru),
- произведений искусств (www.artplanet.com),
- рисунков (www.graphsearch.com),
- книг (<http://bukinist.agava.ru/>),
- кулинарных рецептов (www.cooking.ru/search.html).

В заключение перечислим ПС, предназначенные для поиска файлов:

<http://www.lycos.com/computers/downloads/>, <http://www.filesearch.ru/>,
<http://www.files.ru/>, <http://www.freeware.ru/>.

При использовании перечисленных систем поиск происходит не с помощью ключевых слов, а по известным именам файлов.

Кстати:

Ежедневно поисковая система Google обрабатывает более 5,5 миллиарда запросов (2017 год), причем 15 % поисковых запросов является новыми для поисковой системы» [1].

Лабораторная работа 2. Составление и отладка тематического запроса

Порядок выполнения работы

1. Составить в тетради запрос по теме «Морские сражения во время Великой Отечественной войны». При этом убрать из темы незначащие слова, расширить запрос синонимами, составить логическую формулу запроса с обязательным использованием операторов конъюнкции, дизъюнкции, расстояния и фразы (жесткое словосочетание).

2. Показать запрос преподавателю.

3. Затем записать его варианты на системах Google, Яндекс, Рамблер.

4. Отладить запрос в режиме реального поиска, проводя последовательно сеансы во всех трех системах. Попытаться варьировать поисковые предписания, чтобы добиться оптимальных показателей поиска. Для этого фиксировать в тетради полученные результаты по каждому варианту: точность (по первым 20 документам) и условную полноту (абсолютный объем выдачи).

5. Вернуться к наилучшему поисковому предписанию и текст запроса скопировать через буфер обмена из поисковой строки (окно для ввода запроса)

в окно файла отчета (поочередно в каждой системе). Указать при этом в отчете показатели точности и полноты. Первую веб-страницу с результатами поиска в каждой системе сохранить в своей папке в формате «только html».

6. Найти с помощью системы Яндекс документы, посвященные Льву Гумилеву. Количество найденных документов и сайтов записать в файл отчета. Адрес (URL) первого документа из списка сохранить в Избранном в папке «Гумилев».

7. Затем перейти в режим расширенного поиска и найти документы, посвященные Льву Гумилеву, с датой после 1 октября 2021 г. Новое количество найденных документов и сайтов снова записать в файл отчета. Первый документ из списка результатов поиска сохранить в своей папке в формате «веб-архив, один файл» (*.mht).

8. Найти через систему Рамблер документы по теме «Экономика города Москвы». При этом объем выдачи (количество описаний документов на странице результатов) установить равным 30. Результаты поиска отсортировать по дате (по убыванию) и первую веб-страницу с результатами поиска сохранить в своей папке в формате «только html». Количество найденных документов и сайтов зафиксировать в файле отчета.

9. Перейти в режим расширенного поиска и найти документы по той же теме, но находящиеся лишь на сайте www.rbc.ru. Результаты поиска отсортировать по дате (по возрастанию) и первую веб-страницу с результатами поиска сохранить в своей папке в формате «только html». Количество найденных документов и сайтов зафиксировать в файле отчета.

10. Найти через систему Яндекс документы по теме «Образование», из которых есть ссылка на сайт sutd.ru. Первую веб-страницу с результатами поиска сохранить в своей папке в формате «только html». Количество найденных документов и сайтов зафиксировать в файле отчета.

ТЕМА 3. СПЕЦИАЛЬНЫЕ ВОПРОСЫ ИНФОРМАЦИОННОГО ПОИСКА

Рынок поисковых систем

Google – самая популярная поисковая система в мире с долей на рынке 69,24 %. Bing занимает вторую позицию, его доля 12,26 %.

Таблица 3.1 – Самые популярные поисковые системы в мире

Поисковая система	Доля рынка в июле 2014 г.	Доля рынка в октябре 2014 г.	Доля рынка в сентябре 2017 г.
Google	68,69 %	58,01 %	69,24 %
Bing	17,17 %	29,06 %	12,26 %
Baidu	6,22 %	8,01 %	6,48 %
Yahoo!	6,74 %	4,01 %	5,19 %
AOL	0,13 %	0,21 %	1,11 %
Excite	0,22 %	0,00 %	0,00 %
Ask	0,13 %	0,10 %	0,24 %

В восточноазиатских странах и в России Google – не самая популярная поисковая система. В Китае, например, более популярна поисковая система Soso.

В Южной Корее поисковым порталом собственной разработки Naver пользуется около 70 % жителей. Yahoo! Japan и Yahoo! Taiwan – самые популярные системы для поиска в Японии и Тайване соответственно.

Поисковой системой Яндекс пользуются 53,3 % пользователей в России (Google – 42,9 %).

Большинство «русскоязычных» поисковых систем индексируют и ищут тексты на многих языках – украинском, белорусском, английском, татарском и других. Отличаются они от систем, индексирующих все документы подряд, тем, что, в основном, индексируют ресурсы, расположенные в доменных зонах, где доминирует русский язык, или другими способами ограничивают своих роботов русскоязычными сайтами.

Поисковые системы, учитывающие религиозные запреты

Глобальное распространение Интернета и увеличение популярности электронных устройств в арабском и мусульманском мире, в частности, в странах Ближнего Востока и Индийского субконтинента, способствовало развитию локальных поисковых систем, учитывающих исламские традиции. Такие поисковые системы содержат специальные фильтры, которые помогают пользователям не попадать на запрещённые сайты, например, сайты с порнографией, и позволяют им пользоваться только теми сайтами, содержание которых не противоречит исламской вере.

Незадолго до мусульманского месяца Рамадан, в июле 2013 года, миру был представлен Halalgoogling – система, выдающая пользователям только халяльные, правильные с точки зрения ислама, ссылки, фильтруя результаты поиска, полученные от других поисковых систем, таких как Google и Bing. Двумя годами ранее, в сентябре 2011 года, был запущен поисковый движок I'mHalal, предназначенный для обслуживания пользователей Ближнего Востока. Однако этот поисковый сервис пришлось вскоре закрыть, по сообщению владельца, из-за отсутствия финансирования.

Отсутствие инвестиций и медленный темп распространения технологий в мусульманском мире препятствовали прогрессу и мешали успеху серьёзного исламского поисковика. Очевиден провал огромных инвестиций в веб-проекты мусульманского образа жизни, одним из которых был Muxlim. Он получил миллионы долларов от инвесторов, однако в данный момент этот проект закрыт.

Тем не менее исламские эксперты в области Интернета в течение многих лет занимаются определением того, что соответствует или не соответствует шариату, и классифицируют веб-сайты как «халяль» (хорошо) или «харам» (плохо). Все бывшие и настоящие исламские поисковые системы представляют собой просто специальным образом проиндексированный набор данных либо это главные поисковые системы, такие как Google, Yahoo и Bing, с

определённой системой фильтрации, использующейся для того, чтобы пользователи не могли получить доступ к харам-сайтам, таким как сайты о наготы, азартных играх или каким-либо другим, тематика которых считается антиисламской.

Среди других религиозно-ориентированных поисковых систем распространёнными являются Jewogle – еврейская версия Google и SeekFind.org – христианский сайт, включающий в себя фильтры, оберегающие пользователей от контента, который может подорвать или ослабить их веру.

Пузырь фильтров и поисковая бомба

Пузырь фильтров (англ. «Filter bubble») – негативная сторона персонализированного поиска, явления, при котором веб-сайты определяют, какую информацию пользователь хотел бы увидеть, основываясь на информации о его месторасположении, прошлых нажатиях и перемещениях мыши, предпочтениях и истории поиска. В результате веб-сайты показывают только информацию, которая согласуется с прошлыми точками зрения данного пользователя. Вся иная информация, как правило, пользователю не выводится.

Примером этого являются Google и другие поисковые системы с персонализированными результатами поиска, а также Facebook с персонализированной лентой новостей, которая с каждым действием пользователя наполняется всё более и более персональными результатами. Так, если у пользователя есть два друга в социальной сети, один делает посты с поддержкой руководства, а другой с его критикой, и пользователь делится первыми, а вторые игнорирует, то постепенно в его ленте будут появляться посты со всё более ярко выраженной высокой оценкой руководства. Соответственно, он получает намного меньше противоречащей своей точке зрения информации и становится интеллектуально изолированным в своём собственном информационном пузыре.

Авторами приводится пример, когда один пользователь искал информацию в Google по запросу «British Petroleum» и в ответ получил только инвестиционные новости о компании British Petroleum; когда другой пользователь отправил такой же запрос, то получил в ответ информацию о взрыве нефтяной платформы Deepwater Horizon, и эти страницы с результатами поиска серьезно отличались между собой. Эффект «пузыря» может иметь негативные последствия для формирования гражданского мнения.

Как сказал бывший глава Google Эрик Шмидт: «Людям будет очень сложно увидеть или приобрести что-то, что так или иначе не было под них подобрано».

Кроме того, если операционная и поисковая системы принадлежат одной и той же компании (например, Google и Android, Bing и Windows), то учитывается огромное число персональных данных пользователя: файлы,

приложения; браузер Google Chrome также собирает огромное число персональных данных пользователя.

Исследователи предупреждают, что потенциальным недостатком фильтрации поисковых запросов является то, что она «закрывает нас от новых идей, предметов и важной информации» и «создаёт впечатление того, что наши узкие собственные интересы и есть всё, что существует и окружает нас». Это приносит потенциальный вред, как для личности, так и для общества в целом.

Пузырь фильтров – понятие, похожее на явление, в котором люди и организации ищут информацию, которая изначально кажется им правильной, но она оказывается совершенно бесполезной или почти бесполезной, и избегают информации, кажущейся и воспринимаемой ими как неправильная и несущественная, но оказывающейся полезной.

Проблема появляется потому, что реальная значимость конкретного факта или понятия в этих случаях очевидна только после того, как этот факт стал известен. До этого идея изучения была отклонена, потому что какой-либо факт был неправильно воспринят. Соответственно, ищущий информацию попадает в ловушку и не может узнать то, что ему действительно нужно, попадая в своего рода информационное слепое пятно. Этот феномен был назван «парадоксом релевантности»; он часто встречается на протяжении всего человеческого интеллектуального развития и является важной проблемой для науки и образования.

После публикаций исследования пузыря фильтров, часть пользователей Интернета стала возмущаться тем, что один и тот же, казалось бы, поисковик выдаёт всем различные результаты. Поэтому большинство современных поисковых систем дают возможность регулировки результатов поиска: в Google зарегистрированным пользователям возможно одновременно смотреть персональные и общие результаты с помощью специальной кнопки, находящейся справа от результатов поиска; в Яндексе можно или навсегда включить, или навсегда отключить персональный поиск; в Bing можно пока только очистить историю просмотра.

Поисковая система DuckDuckGo вообще не использует пузырь фильтров и не собирает информацию о пользователях. Она позиционирует себя как использующую информацию из многих источников с целью предоставления более точных, актуальных и разнообразных результатов поиска, чем обычные поисковики.

Поисковая бомба (англ. linkbomb) – феномен функционирования поисковой системы, когда какой-нибудь запрос приводит к выдаче абсурдного или провокационного результата. Поисковые бомбы иногда являются случайной ошибкой поисковой системы, но многие из них, особенно связанные с острыми социальными и политическими конфликтами, провоцируются искусственно. Для этого используется техника поискового спама. Релевантность поисковой выдачи искажается из-за прописывания ссылок с ключевыми словами. Так, на запрос More Evil Than Satan Himself («большее зло, чем сам сатана») в начале 2000-х годов первой ссылкой появлялся Билл Гейтс,

Miserable failure («досадная ошибка») – президент США Джордж Буш-младший, и таких примеров достаточно. В любых поисковых системах есть алгоритмы, позволяющие обнаруживать «поисковые бомбы» и фильтровать результаты, возникающие при использовании этого феномена.

Лабораторная работа 3. Использование пользовательского поиска в Internet

Порядок выполнения работы

На основе поисковых систем нового поколения можно строить свои собственные индивидуальные или коллективные поисковые машины. Поиск можно адаптировать к определённой тематике и к определённому сообществу. Пример организации социального поиска – персональная или групповая поисковая машина от Google находится по адресу: <http://www.google.com/coop/cse/>

Создание собственного поисковика в **Google CSE** начинается с определения списка сайтов, которые пользователь считает объективными и актуальными источниками информации в различных сферах деятельности. К работе по изменению созданной поисковой системы можно подключить других пользователей, которые могут стать соавторами, и настраивать ее свойства коллективно. Домашняя страница созданной поисковой машины, помимо строки поиска, будет включать профиль создателя и список заданных сайтов.

1. Создайте систему пользовательского поиска в **Google CSE**. Определите список сайтов, которые будут включены в систему пользовательского поиска. Перейдите на страницу системы персонального поиска **Google CSE** – <http://www.google.com/coop/cse/> . Если есть e-мейл на google.com – введите его для перехода на эту страницу. Если такого e-мейла нет, его нужно создать.

2. Нажмите кнопку *Создать систему персонального поиска*. Дайте название системе. Опишите систему (для чего она создается, что можно будет находить при помощи этой поисковой машины, среди каких сайтов будет вестись поиск и т.п.). Поищите ключевые слова для поиска системы в сети.

3. Добавлять сайты к системе можно на странице управление.

Панель управления: Поиск по образовательным Вики

Основные сведения | [Сайты](#) | [Уточнения](#) | [Внешний вид](#) | [Код](#) | [Совместная работа](#) | [Заработайте!](#) | [Версия для компаний](#) |

Основные сведения

Название и описание будут показываться на [домашней странице](#) Вашей поисковой машины в Google.

Название системы поиска:

Описание системы поиска:

Ключевые слова, которые описывают тему или содержимое Вашей поисковой машины. Эти ключевые слова используют Вашу поисковую машину. [Узнать больше](#).

Ключевые слова системы поиска:
например: климат "глобальное потепление" "парниковые газы"/>

Язык системы поиска:

Кодировка системы поиска:

Рисунок 1. Описание персональной системы поиска

К работе по изменению созданной поисковой системы можно подключить других пользователей. Вы можете разрешить им стать соавторами поисковой системы и настраивать ее свойства коллективно.

Добавить
аннотацию в:

[изменить систему поиска »](#)

☒ Включить ☐ Исключить

Включить:

☒ Найти на этом сайте (www.ourstories.org)
☐ Только эта страница (www.ourstories.org/ourpartners.html)

Применить ярлыки:
(необязательно)

☒ примеры Гугл ☒ младшие школьники ☒ зарубежный опыт ☐ наука

Рисунок 2. Добавление сайта к системе пользовательского поиска

Для создания пользовательских систем необходимо ввести ряд ключевых слов согласно следующей таблице:

Таблица 3.2 – Варианты заданий лабораторной работы № 3

Вариант	Ключевые слова
1	Информатика информационные технологии компьютерная графика поиск информации
2	Физика ядерная физика физика элементарных частиц бозон
3	Химия химические реакции неорганическая химия щелочные растворы
4	Лимонад безалкогольные напитки прохладительные напитки буратино
5	Музей бесплатный вход музеи Петербурга эрмитаж
6	Аэропорт билет в Москву Пулково самолет
7	Телевидение телепрограмма канал СТС телесериал
8	Трамвай общественный транспорт рельсовый транспорт маршрут 16
9	Мороженое сладости пломбир крем-брюле
10	Российская империя императоры Петр Первый зимний дворец
11	Метрополитен петербургское метро станция закрытого типа метродепо
12	Лес зимний лес леса России смешанный лес
13	Железная дорога детская железная дорога лего железная дорога поезд
14	Футбол турнирная таблица чемпионат России РФПЛ
15	Александр Пушкин тексты биография няня

После создания пользовательской системы необходимо осуществить поиск с каждым из ключевых слов и/или сочетаний слов, записать в отчёте статистику по каждому запросу. Сделать вывод о том, насколько удачная пользовательская система была создана. После чего произвести те же операции в обычной поисковой системе без установки пользовательских ограничений.

ТЕМА 4. АНАЛИЗ УДОБСТВА (USABILITY) ПОИСКА ДЛЯ ИНФОРМАЦИОННОГО САЙТА

«Коммерческий интернет-сайт должен быть эффективным бизнес-инструментом, а не визитной карточкой фирмы. Сайт, который создан для того, чтобы продавать, должен именно эффективно продавать товары или услуги фирмы, а значит, быть максимально удобным в использовании. Сайт должен быть сделан именно для пользователя, а не для сотрудников фирмы и генерального директора.

Юзабилити (usability) – это удобство использования чего-либо. В нашем случае под этим термином подразумевается удобство использования сайта. Сайт должен быть максимально удобным и удовлетворять определенным общепринятым стандартам и нормам юзабилити, которые выработаны сообществом пользователей.

Существует множество определений юзабилити. Среди них можно выделить следующие:

- согласно стандарту ISO-13407 (User Centered Design, UCD): «Ориентированный на пользователя дизайн характеризуется активным вовлечением пользователя в процесс разработки для достижения прозрачного понимания пользовательских требований и надлежащего распределения функций между пользователями и технологиями, а также итеративным характером подхода и его мультидисциплинарностью».

- согласно стандарту ISO 9241-11: «Степень, с которой продукт может быть использован определенными пользователями при определенном контексте использования для достижения определенных целей с должной эффективностью, продуктивностью и удовлетворенностью».

Некоторые стандарты юзабилити, которые могут быть применены для большинства коммерческих сайтов, приведены в следующих таблицах.

Таблица 4.1 – Общие требования к сайту

№	Параметр	Значение
1	Время загрузки	Не более 10 секунд при скорости 26 400 Кбит/с (модем). Размер файла страницы не более 30 — 70 Кбайт
2	Браузеры	Сайт должен одинаково просматриваться в браузерах Internet Explorer (IE) и Mozilla двухгодичной давности. Сайт также должен отображаться без грубых погрешностей в других браузерах: Opera, Netscape и т. п.
3	Разрешение монитора	Сайт должен хорошо просматриваться при разрешении от 800 × 600 пикселей
4	Обработка ошибки 404	Страница-заглушка, сообщающая о том, что данного документа не существует, и рекомендуемая воспользоваться поиском или перейти на главную страницу сайта
5	Реклама	Указание на то, что область является рекламой
6	Технические проблемы	Страницы сообщения о технических проблемах с указанием контактной информации и времени устранения проблем и неполадок
7	Авторизация пользователей	Область авторизации должна быть расположена в заметном месте (без прокрутки страницы при разрешении 800 × 600 пикселей)

Таблица 4.2 – Структура сайта

№	Параметр	Значение
1	Ширина страницы	Оптимизированная для отображения в окне шириной 770 пикселей, с гибкой структурой страницы, которая подходит для любого окна размером от 620 до 1024 пикселей
2	Структура страницы	Гибкая, автоматически приспосабливается под конкретный размер окна пользователя
3	Длина страницы	1—2 HTML – экрана ¹ . Не более трех HTML – экранов (от 1000 до 1600 пикселей). Необходимо, чтобы на одном экране уместились все главные элементы страницы
4	Использование фреймов	Нет
5	Иерархия разделов	Принцип «трех кликов». Трехуровневая структура сайта, чтобы пользователь мог попасть в любой раздел сайта за три щелчка
6	Горизонтальная прокрутка	Ни в коем случае
7	Вертикальная прокрутка	Стараться избегать
8	Раскрывающиеся окна (popup)	Не использовать

Таблица 4.3 – Требования по навигации сайта

№	Параметр	Значение
Меню		
1	Основное меню	Слева и сверху, должно быть в заметном месте. Ссылки должны быть текстовыми, а не графическими
2	Дополнительное меню	Дублирование ссылок на разделы сайта внизу, если страница имеет более одного HTML – экрана
3	Карта сайта	Обязательно, со ссылками на все разделы сайта. Для сайтов, где количество страниц невелико, можно не использовать. Для больших сайтов делать карту в виде иерархического дерева

4	Иерархическое меню / «хлебные крошки»	Допускается наличие дополнительной навигационной строки, показывающей иерархию разделов и местоположение человека. В качестве разделителя желательно использовать знак «>». Пример: Главная > Отели Турции > Отель Wow Торкаррі
Поиск		
1	Наличие поиска	Очень желательно
2	Расположение	Вверху сайта. Оптимально — правый верхний угол. Или левый верхний угол
3	Поле поиска	Цвет поля — белый. Ширина — примерно 30 символов
4	Название кнопки поиска	Найти
5	Расширенный поиск	При необходимости. Должен быть вынесен на отдельную страницу. Должна быть ссылка с основного поиска
6	Область поиска	По умолчанию ищет только по сайту
7	Если ничего не найдено	Предложение расширенного поиска или сообщение о том, что по запросу ничего не найдено, с предложением переформулировать запрос или обратиться в офис компании
Ссылки		
1	Self – ссылки	Не должно быть ссылок с текущей страницы на саму себя
2	Названия ссылок	Логичные и понятные, отличные друг от друга
3	Цвета ссылок	Обязательно должны различаться цвета просмотренных и не просмотренных ссылок. Не просмотренные ссылки выделяются голубым, просмотренные — иным контрастным цветом (или по усмотрению дизайнера)
4	Функциональные ссылки	Если идет ссылка на скачивание файла / анимации / музыки, то должно быть указание на это в виде объема данных и типа носителя (файл, музыка, видео). В отдельных случаях можно указывать время скачивания, которое потребуется при заданной пропускной способности канала
5	Вид ссылок	Ссылки должны быть подчеркнуты (за исключением ссылок на панели навигации)
6	Тип ссылок	Текстовые ссылки, не графические

Таблица 4.4 –Требования по графическому шрифтовому оформлению сайта

№	Параметр	Значение
1	Объем графики	Не перегружать страницу графикой. Графика не должна превышать 30—50 Кбайт на страницу. Исключение: страницы, специально предназначенные для демонстрации графики: фотоальбом и т.п.
2	Подписи рисунков (alt)	Все рисунки должны быть понятно подписаны (тег alt), чтобы с выключенной графикой сайт нормально отображался и понимался пользователем. В тегах обязательно задавать атрибуты height и width
3	Флеш–анимация	Должна быть использована по минимуму либо по ссылке на отдельной странице
4	Файлы мультимедиа	При размещении на сайте аудио, видео и иной мультимедийной информации должны быть учтены следующие требования. 1. Указаны формат и объем мультимедийного файла. 2. Указано время загрузки мультимедийного файла (для файлов объемом более 50 Кбайт). 3. Сделан образец для предварительного просмотра (прослушивания) мультимедийной информации. Для фильмов – 1 – 2 кадра, для звуковых файлов – фрагмент. 4. Должны использоваться форматы кодирования мультимедийной информации, поддерживаемые старыми версиями браузеров
5	Фон и текст (шрифт)	Существуют следующие требования. 1. Фон и текст страницы должны быть оформлены контрастными цветами. 2. Наилучшее сочетание: черный текст на белом фоне (белый текст на черном фоне). В качестве фона страницы должен быть использован однотонный цвет либо узор с минимальной контрастностью. 3. Текст должен быть набран достаточно крупным шрифтом. 4. Текст должен быть статичным. 5. Желательно выравнивание текста по левому краю
6	Размер шрифта	Желательно 12 пунктов*
7	Вид шрифта	Желательно шрифт без засечек Verdana или Arial*
8	Основной текст	Желательно черный*
9	Фон	Желательно белый*
10	Расположение логотипа	Верхний левый угол страницы

11	Размер логотипа	От 58 × 58 до 111 × 111 пикселей (оптимально 80 × 68 пикселей)*
12	Ссылка с логотипа	Обязательно на главную страницу
13	Заставка при загрузке сайга	Использование нежелательно

Аудит юзабилити

Аудит юзабилити направлен на то, чтобы, с одной стороны, оценить текущее состояние удобства сайта, а с другой стороны, путем анализа выявить недостатки юзабилити и улучшить удобство сайта.

При этом аудит юзабилити можно разделить на два вида:

1) тот, который производится специалистом организации по стандартам и исходя из собственного опыта и знаний;

2) тот, который выполняется сторонними пользователями (пользовательский аудит юзабилити).

1. Аудит юзабилити специалистом организации. В процессе анализа сайта должны быть оценены определенные части сайта и по каждой части выполнен анализ ошибок и выданы подробные рекомендации. Анализируются:

- главная страница сайта;
- навигация по сайту;
- структура сайта;
- графическое и шрифтовое оформление сайта;
- информационное наполнение сайта.

Целесообразность проведения аудита других частей сайта определяется специалистом самостоятельно исходя из специфики задачи и самого сайта.

Кроме того, может производиться аудит юзабилити следующих частей сайта:

- отдельных значимых разделов сайта (каталога товаров и услуг, форума, базы данных и т. п.);
- тематических рассылок, которые осуществляются с сайта;
- администраторской системы сайта;
- форм обратной связи (анкеты, форумы, гостевые книги).

2. Аудит юзабилити пользователями. Преимущества аудита пользователями состоят в том, что вы видите, как с вашим сайтом работают живые люди, с какими трудностями они сталкиваются, что у них получается, а что не очень. Кроме того, такой аудит довольно легко проводить и без привлечения сторонних подрядчиков. Достаточно сформулировать цели, задачи (список вопросов), принести видеокамеру и подобрать от трех до пяти «подопытных кроликов». В пользовательском аудите возможны три варианта.

Вариант 1 – анкетирование. Пользовательский аудит юзабилити производится путем анкетирования.

Для сайта разрабатывается специальная анкета, содержащая вопросы по сайту и предлагающая пользователям решение конкретных задач на сайте. В анкете фиксируются мысли и мнения пользователей.

Пользовательский аудит может проходить поэтапно (три этапа) либо в один этап. Для аудита достаточно привлечь 4–5 пользователей, причем не обязательно из целевой аудитории. Очень желательно, чтобы пользователи,

привлекаемые для аудита сайта, имели различный опыт использования Интернета. В ряде случаев могут быть определены и иные ограничения для участников тестирования.

Специалист, осуществляющий аудит сайта, разрабатывает анкету и рассылает ее участникам тестов. После заполнения анкет специалист анализирует выводы и мнения пользователей и составляет подробные рекомендации по улучшению сайта, основываясь на выводах пользователей и личном опыте.

Вариант 2 – тестирование «вживую». Аудит юзабилити может проводиться непосредственно пользователями, при этом применяются средства наблюдения (видеокамера, специальные программы, позволяющие отслеживать перемещение взгляда пользователя).

Наиболее доступным является следующий *Вариант 3 аудита:* пользователь располагается перед монитором и выполняет различные задания, комментируя свои действия. Все происходящее фиксируется на видео (можно также параллельно записывать мысли пользователя на бумагу). Затем собранная информация обрабатывается и формируется отчет об ошибках» [2].

Лабораторная работа 4. Оценка качества Интернет-ресурса

Порядок выполнения работы

Протестировать предложенные веб-страницы:

<http://ngrp.telefun.ru/>

<http://www.budgetrf.ru>

<http://www.edu.ru>

<http://www.sci.aha.ru/>

<http://www.nstu.ru>

<http://gov.ru>

<http://www.ru.emb-japan.go.jp/>

<http://www.spsl.nsc.ru/>

<http://www.tomsk.ru/>

<http://www.ras.ru/>

<http://www.hse.ru/>

<http://www.stanford.edu/>

<http://www.cfin.ru/>

<http://www.berkeley.edu/>

<http://www.philosophy.ru/>

Порядок выполнения:

1. Зайти на страницы предложенных веб-ресурсов, оценить качество каждого ресурса.

2. Заполнить таблицу «Качество веб-ресурса». Каждый пронумерованный критерий оценивается в 10 баллов.

3. Таблицу в отчете расположить горизонтально, рассчитать для каждого ресурса итоговый рейтинг.

4. Для одного из выбранных ресурсов (по желанию) в отчете описать обоснование проставленных баллов.

5. По итогам выполнения работы представить список сайтов, начинающийся с самого качественного, согласно выполненным расчётам, и заканчивающийся наименее удачным Интернет-ресурсом.

Таблица 4.5 – Качество веб-ресурса

1. Достоверность	Автор (ответственное лицо) и возможность связи	
	Автор (ответственный) адекватен данному ресурсу?	
2. Цель	Заявлена цель ресурса. Она соответствует вашим ожиданиям	
3. Точность	Имеется возможность связи с автором (ответственным лицом): электронная почта или контактный адрес /	
4. Управление	Автор (ответственный) и веб-мастер разделены	
	Домен. Регистрация домена ресурса адекватна его целям и задачам	
5. Авторитетность	Чем подтверждена авторитетность авторов (ответственных) ресурса	
6. Объективность	Имеются ли ссылки на альтернативные мнения (источники)	
	Насколько подробной является информация	
	Выражены ли альтернативные мнения	
	Не является ли данный ресурс скрытой рекламой	
7. Оперативность	Дата создания	
	Дата последнего обновления	
	Имеются ли актуальные материалы (ссылки на таковые)	
8. Актуальность	Является ли информация на странице устарелой?	
	Сколько «мертвых» ссылок на странице?	
9. Удобство	Для просмотра одной порции информации хватает 1-2-х экранов	
	Имеются ссылки (если есть), дополняющие ресурс, расширяющие информационное поле	
	Имеется карта сайта	
	Соблюден баланс текста и иллюстраций	
	Не требуются дополнительные программы и модули для просмотра данных	
10. Доступность	Бесплатность доступа	
	Необходимость дополнительных регистраций	
	Есть ли проблемы при просмотре ресурса с тем браузером, которым вы пользуетесь	
Сочетание всех параметров	Рейтинг по 100-балльной системе	

ТЕМА 5. АНАЛИЗ СПРОСА НА ТОВАРЫ САЙТА СРЕДИ АУДИТОРИИ ПОИСКОВЫХ СИСТЕМ

Спрос в поисковых системах может быть легко изучен с помощью статистики запросов поисковых систем. Суть статистики запросов в том, что можно посмотреть, сколько раз в предыдущем месяце пользователи поисковых систем делали тот или иной запрос, а также найти сходные запросы, с помощью которых могут искать ваши товары или услуги. На языке профессионалов процесс поиска запросов и их группировки называется составлением семантического ядра сайта.

«Семантическое ядро является фундаментом продвижения любого сайта. Дадим определение этому понятию. *Семантическое ядро запросов сайта* – это набор запросов, которые относительно часто используются пользователями при поиске информации, имеющейся на сайте или близкой к ней по теме.

Когда пользователи хотят найти информацию, они редко используют научные формулировки, поэтому нельзя заранее предсказать, как будет сформулирован их запрос. Чтобы проводить оптимизацию сайта для поисковых машин и понимать, какими словами обозначают свою потребность пользователи (что можно использовать в рекламе), необходимо найти соответствие между темой сайта и набором формулировок запросов. Этот набор и является семантическим ядром запросов.

При составлении ядра принимается во внимание несколько факторов.

а) Учитывается смысл запроса, а не его форма. Даже если поисковая фраза содержит «правильные» слова, она может быть нецелевой (например, используя запрос «телефоны Москвы», ищут, очевидно, справочник телефонов, и для продажи телефонов он не годится). Ориентация на смысл требует ручного труда, что делает составление семантического ядра очень трудоемким.

б) Учитываются точные формулировки запросов. Грамотный подход к формированию ядра требует учета точных формулировок, потому что современные поисковые системы даже при перестановке слов выдают другие результаты, не говоря уже о добавлении уточняющих слов в запрос. Даже если сайт будет на первом месте по запросу «мобильные телефоны», это не значит, что он попадет в первую сотню по запросу «мобильные телефоны Siemens», сколько бы раз слово Siemens ни встречалось на странице. Это еще одна причина, существенно усложняющая составление семантического ядра.

в) Для включения в ядро запрос должен быть достаточно частотным, иначе он может оказаться случайным, а поскольку таких запросов очень много, время будет потрачено зря. Из этого правила есть два исключения. Первое – группа очень дорогих товаров, когда каждый случайный посетитель может принести большую прибыль. Второе исключение – специальные методы раскрутки, когда группа низкочастотных запросов помещается на страницу в виде списка; в этом случае очень важно продумать содержание и внешний вид страницы, чтобы она не оттолкнула потенциального клиента и не испортила впечатление о компании.

г) Учитывается многозначность запроса. Если пользователь задал запрос «аквариум», неизвестно, хочет ли он узнать что-либо об аквариуме для своих рыбок или интересуется группой «Аквариум». Решение о включении запроса принимается на основании предположений о доле целевых посетителей среди тех, кто задал данный запрос.

д) Исключаются маловероятные и слишком узкие запросы. Если на сайте предлагается широкий перечень товаров, исследование всех запросов становится нереальным. В таком случае, если спрос в исследуемой области достаточно большой, из ядра исключаются недорогие товары, по которым маловероятен запрос.

В отличие от анализа посещаемости и поведения посетителей на сайте, семантическое ядро позволяет оценить внешний по отношению к сайту спрос, изучить потенциальную аудиторию, которая еще не пришла на сайт.

Полученные формулировки запросов можно использовать в следующих целях:

- 1) для оптимизации сайта для поисковых систем – запросы применяются для формирования текста страниц и ссылок; для контекстной рекламы – ключевые слова для показа объявлений выбираются среди ядра запросов; эти же поисковые фразы включаются в текст объявлений;
- 2) для обычной рекламы – если реклама содержит текст или слоган, желательно, чтобы он включал самую часто встречающуюся поисковую фразу из ядра, подходящую по теме;
- 3) для создания фильтров и сортировок на сайте;
- 4) для выдачи задания копирайтеру на написание контента;
- 5) для проектирования нового сайта или переделки структуры текущего сайта;
- 6) при открытии новых направлений в бизнесе; для понимания сезонности спроса.

Семантическое ядро – это база для всего: построения структуры сайта, написания текстов, создания контекстных рекламных кампаний. Если будет плохо проработана база, такими же плохими будут результаты от продвижения.

Анализ спроса можно разбить на несколько этапов, главным из которых является сбор поисковых запросов.

Для получения семантического ядра запросов сайта нужно выполнить несколько действий.

1. Анализ тематической области и текстов сайта с целью формирования начального списка запросов.
2. Поиск возможных синонимов и однокоренных слов для всех ключевых терминов (для поисковых систем «компьютер» и «компьютерный» – два разных слова).
3. Поиск сходных по смыслу запросов.
4. Поиск сопутствующих терминов, которые употребляются только в контексте целевой тематики. Например, запрос «3310», без слова «nokia»

составляет примерно 12 % от частоты запроса «nokia 3310», причем 90 % вторичных поисковых запросов относятся именно к телефону Nokia 3310.

5. Выявление терминов, которые используются как в русском, так и в английском языке.

6. Анализ популярных запросов с целью выявления вариантов транслитерации (например, «сименс» и «сиеменс»), жаргонных названий («мобильный телефон» и «мобила»), орфографических ошибок («ericsson» и «ericson»).

7. Анализ частот выявленных запросов с помощью инструментов статистики «Рамблера» (<http://adstat.rambler.ru/wrds/>) и «Яндекс.Директ» (<http://direct.yandex.ru>).

8. Исключение из списка нецелевых, редких, а также многозначных запросов. Для определения того, насколько запрос целевой, применяется человеческий интеллект, а также инструменты статистики» [2].

Лабораторная работа 5. Составление запросов на поиск информации в сети Internet

Порядок выполнения работы

УПРАЖНЕНИЕ 1.

Изучите назначение, интерфейс и возможности поисковых серверов WWW – по выбору: <http://www.yandex.ru>, www.google.ru или какой-то другой по своему выбору. В отчете написать краткую информацию про эти поисковые серверы (со скриншотами). Изучите язык запросов одного из поисковых серверов WWW.

УПРАЖНЕНИЕ 2.

Найдите в Интернете источники информации о вашем любимом музыкальном исполнителе. Если он иностранец, то попробуйте найти иноязычные источники информации. Используйте не менее 4 источников. Запишите адреса найденных вами страниц.

Отчет создать в виде:

Исполнитель (или группа) _____

Сайты с информацией об исполнителе (или группе) _____

Информация _____

УПРАЖНЕНИЕ 3.

Используя возможности поиска информации в Internet, найдите фирмы, которые продают специализированные образовательные пакеты (программы) или предоставляют услуги для работы с ними.

В отчет вставить следующую таблицу и заполнить:

№	Фирма	Программа	Стоимость программы	Дополнительная информация
1				
2				
3				

УПРАЖНЕНИЕ 4.

Найти ответы на вопросы, используя «точные» запросы и язык запросов одного из вышеуказанных поисковых серверов.

1. Место и дата рождения Президента России. Найти его фотографии и отрывки из последних выступлений или интервью и сохранить их в вашем пользовательском каталоге.

2. Место и дата рождения теннисиста Евгения Кафельникова. Найти его фотографии и отрывки из последних интервью и сохранить их в вашем пользовательском каталоге.

3. Фамилию первого тренера теннисиста Марата Сафина. Найти фотографии теннисиста и отрывки из последних интервью и сохранить их в вашем пользовательском каталоге.

4. Когда и где родился А. И. Солженицын. Составить список его произведений. Найти его фотографии, сохранить их в вашем пользовательском каталоге.

5. Когда и где родился английский писатель Дж. Р. Р. Толкиен. Составить список его произведений. Найти его фотографии, сохранить их в вашем пользовательском каталоге.

6. Назвать режиссера фильма «Титаник». Когда и где он родился. Найти его фотографии, сохранить их в вашем пользовательском каталоге.

7. Когда и где родился В. И. Даль. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

8. Когда и где родился Ф. А. Искандер. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

9. Когда и где родился В. С. Высоцкий. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

10. Когда и где родился Л. Н. Толстой. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

11. Когда и где родился В. В. Гиляровский. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

12. Когда и где родился В. Войнович. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

13. Когда и где родился А. Г. Битов. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

14. Когда и где родилась И. Хмелевская. Составить список ее произведений. Найти ее фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

15. Когда и где родился Л. Н. Гумилев. Составить список его произведений. Найти его фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

16. Когда и где родились А. Н. и Б. Н. Стругацкие. Составить список их произведений. Найти их фотографии или отрывки из произведений и сохранить их в вашем пользовательском каталоге.

УПРАЖНЕНИЕ 5.

Используя сведения из сети Internet, подготовить тезисы доклада по указанной теме. Составьте список источников, из которых вы берете информацию для доклада (по вариантам).

1. Приведение региональных законодательных актов в соответствии с основным законом РФ.

2. Информационные модели.

3. Основы языка HTML и структура HTML – документа.

4. Предмет «информатика» и его связь с естественными и гуманитарными науками.

5. Принципы фон Неймана работы ЭВМ.

6. История возникновения сети Internet.

7. Информационные технологии: возможности и ограничения.

8. Краткая история вычислительной техники.

9. Основы технологии «клиент – сервер».

10. Основы компьютерных коммуникаций

11. Классификация персональных компьютеров.

12. Классификация информационных систем.

13. Вычислительные сети: основные возможности и проблемы работы в сетях.

14. Программы-переводчики: основные характеристики и возможности.

ТЕМА 6 . КЛАССИФИКАЦИОННЫЕ СИСТЕМЫ ПОИСКА ДОКУМЕНТОВ

Все разнообразие классификационных систем поиска документов представлено в следующей схеме.

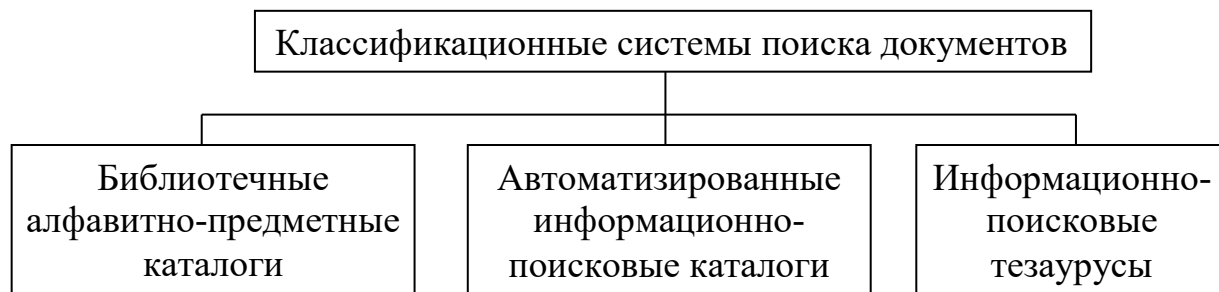


Рисунок 4. Классификационные системы поиска документов

Современные библиотечные классификации основываются на системах десятичной классификации Мелвила Дьюи и правилах построения алфавитно-предметных рубрик Ч. А. Каттера.

«Для поиска в классификационных системах поиска документов осуществляется классификационное индексирование документов. Виды индексирования представлены на рисунке.



Рисунок 5. Виды индексирования документов

Перечислительная классификация

Формируется список предметных заголовков, расположенных в алфавитном порядке. Каждая предметная рубрика получает определенный цифровой

или буквенный код. Документ индексируется перечислением кодов рубрик, отражающих его содержание (предметы). Особенностью систем перечислительной классификации является то, что документ может быть индексирован любым количеством предметов (рубрик), отражающих его содержание.

Для поиска по классификатору (каталогу) определяются коды интересующих пользователя предметов (рубрик), и далее отбираются из хранилища документы, проиндексированные соответствующими кодами. Для удобства поиска и отбора по каждому документу формируется специальная карточка, на которой содержание документа отражается как кодом, так и рефератом на естественном языке. Кроме того, на карточку обычно заносится информация о его физическом местонахождении, об авторе, названии и другие библиографические данные документа. Поиск и отбор осуществляется по карточкам, с последующим извлечением из хранилища самих документов.

Основным *недостатком* перечислительной классификации является отсутствие систематизированных связей и отношений между предметными рубриками.

В определенной степени этот недостаток преодолевается использованием перекрестных ссылок через конструкцию "см. также", ориентирующих пользователя на смысловую связь некоторых рубрик.

Иерархическая классификация

В данном виде классификации список предметных рубрик строится как иерархическая структура в виде перевернутого дерева. Вся предметная область разбивается на ряд взаимоисключающих (непересекающихся) рубрик, которые, в свою очередь, могут включать несколько подрубрик по принципу "Род-вид". Таким образом уже используются некоторые семантические основы предметной области.

Содержание документа индексируется кодами соответствующих рубрик, однако при этом отпадает необходимость в явном указании более общих рубрик, если в индексе есть код их подрубрик.

Иерархический и перечислительный подходы к классификации впоследствии воплотились в широко используемых в библиотечной практике алфавитно-предметных каталогах, наиболее распространенным из которых в настоящее время является универсальная десятичная классификация (УДК), в основе которой лежит классификационная схема Дьюи, дополненная правилами образования сложных рубрик, а также специальными определителями для более детального описания документов (определители формы и характера документа, определители времени и т.д.). Систематизированная классификация позволяет строить каталог или картотеку в структурно-иерархическом виде, повышающем точность поиска.

Недостатком перечислительной и иерархической классификации является принципиальная невозможность заранее перечислить все темы, а добавление к классификатору новых рубрик требует переиндексирования всего ранее накопленного фонда документов.

Фасетная классификация

Предложена в 1930-х годах индийским библиотековедом и математиком Шиали Рамамрита Ранганатаном.

Идея фасетной классификации, преодолевающей в определенной степени ограничения двух предыдущих классификаций, состоит в том, что вся предметная область сведений разбивается на ряд исходных групп – фасет по организационно-технологическому или семантическому принципу, отражающему специфику предметной области.

Фасеты выступают в роли «кирпичей», из которых можно сконструировать любую, даже самую сложную и узкую предметную область. Внутри фасет предметные рубрики строятся и упорядочиваются по алфавитно-иерархическому принципу.

Пример. Предметная область документов по производству технологического оборудования разбивается на 4 фасета: *Конфигурация, Материалы, Напряжения и нагрузки, Типы разрушений*.

Как видно из рисунка, документ о нагрузках на сжатие трубчатых конструкций из никелевых сплавов получит индекс AcBgtLg.

АКонфигурация Ac Трубчатые АКонфигурация	ВМатериалы Bc Металлы BccЖелезные сплавы BcdСталь и BcfaХромистые стали BcfiХромоникелевые стали BgtНикелевые сплавы
К Типы разрушений Kg Ползучесть KgbДлительная прочность KiХрупкий излом Kl Продольный изгиб Kp Усталость Kpf Коррозийная усталость	L Напряжения и нагрузки Lb Растяжение LdКручение LeСрез LgСжатие Lm Удар

Рисунок 6. Пример фасетной классификации

Основное достоинство: небольшим перечнем фасетных рубрик возможно сконструировать огромное количество узких специализированных рубрик, точно и полно проиндексировав содержание документа.

Недостаток: влияние на эффективность поиска документов порядка следования рубрик фасет.

Психологически пользователь при поиске в первую очередь сосредотачивает внимание на обозначениях подрубрик, стоящих первыми в цепном списке сконструированной формулы. Если интересующие его сведения отражаются рубрикой, стоящей не в начале списка, то он может «с ходу» отвергнуть всю формулу. Для преодоления этого недостатка используется *пермутация*, при которой для документа приводится список всех возможных вариантов написания сконструированной фасетной формулы на основе циклической перестановки.

Например: AcBgtLg
LgAcBgt
BgtLgAc и т.д.

Такой подход не всегда решает проблему, поскольку число перестановок может быть велико, а это усложняет поиск.

Другим подходом, напротив, является жесткая регламентация порядка изложения фасет, ориентируя первоначальное внимание пользователя на интересующих его в первую очередь фасетах.

Сильной стороной фасетной классификации является более глубокое, чем при иерархической классификации, использование семантики.

Координация понятий в классификационных системах

Координация понятий выражается в использовании различных операций над совокупностью понятий при индексировании документов или при поиске документов. Выделяются два направления – использование только логических операций и использование определенной грамматики понятий, классов, рубрик в рамках определенного семантического языка.

Под *классом* (понятием) будем понимать совокупность (множество) документов, проиндексированных кодом соответствующего класса.

Логические операции над понятиями дают возможность отображения при индексировании документов и формировании запросов более сложных и многоаспектных понятий.

Пример. Рубрика AcBgtLg из примера по фасетной классификации в терминах логических операций представляет собой пересечение трех классов $Ac \cap Bgt \cap Lg$.

При этом использование дополнительных операций объединения (ИЛИ) и отрицания (НЕ) обеспечивает построение более сложных, чем при чисто фасетной классификации, комбинаций рубрик, классов и понятий. А операция включения дает возможность так называемого цепного представления и описания иерархических структур каталогов, формализуя систематизированный аспект классификаторов.

Предкоординация понятий предусматривает использование операций над классами *при индексировании документов*. Иначе говоря, индекс документа представляет собой конструкцию из исходных понятий (классов) классификатора, построенную на основе логических операций.

Посткоординация понятий предусматривает осуществление операций над классами *при поиске документов*, т.е. в процессе формирования поискового образа запроса. Технология и механизм поиска при этом включает предварительный отбор всех документов с индексами классов (рубрик), входящих в логическую конструкцию запроса, с последующим проведением собственно логических операций над отобранными множествами документов.

Информационно-поисковые тезаурусы

Большую роль в развитии информационно-поисковых систем сыграла система унитермов, согласно которой содержание документа индексируется совокупностью терминов в виде *однословных обозначений – унитермов*, в качестве которых чаще всего выступают элементы словаря *ключевых терминов* по определенной предметной области.

Пример. Документ по теории информационного поиска может быть проиндексирован унитермами – «Информационный» и «Поиск».

Недостатки системы унитермов, приводящие к неоднозначности индексирования документов:

- *ложная координация*, заключающаяся в том, что определенные понятия присутствуют в документе, но комбинируются по смыслу с другими понятиями (например, документ об информационном обеспечении поисковых бригад при ликвидации чрезвычайных ситуаций содержит и унитермы "Информационный" и "Поиск", а следовательно, неправильно может быть выдан на запрос по теории информационного поиска;

- *проблема синонимичности и омонимичности*, заключающаяся в наличии в текстах синонимов (одинаковых или близких по смыслу слов) и омонимов (одинаковых в написании и звучании слов, но имеющих разный смысл – рукав одежды, рукав реки).

Для преодоления этих недостатков в чисто дескрипторную систему унитермов ввели *составные термины, указатели связи и ролей терминов*, внося тем самым элементы семантики. Это привело к созданию информационно-поисковых тезаурусов.

Тезаурус (с греч. «хранилище», «запас», «сокровищница») в узком смысле представляет собой специальный словарь-справочник, в котором перечислены ключевые слова-дескрипторы определенной предметной области, установлены способы устранения синонимии, омонимии, полисемии, определены родовые и ассоциативные связи дескрипторов.

В отличие от информационно-поисковых каталогов в тезаурусах помимо классификационной схемы присутствуют сами дескрипторы – ключевые слова.

Главная идея информационно-поисковых тезаурусов заключается в повышении эффективности и автоматизации индексирования в рамках дескрипторного подхода. Т.е. поисковый образ документа задан набором дескрипторов, но при его индексировании учитываются семантические отношения между дескрипторами.

Достоинства тезаурусов:

- обеспечивают более адекватный содержанию поисковый образ документа;

- повышают эффективность поиска документов по точности, полноте и шуму.

Пример. Согласно ГОСТ 18383-73 форма представления тезауруса включает алфавитное перечисление статей по каждому дескриптору (термину) в следующем виде:

...

РЕФЕРАТ

с резюме
в СВЕРТЫВАНИЕ ИНФОРМАЦИИ
н РЕФЕРАТ АВТОРСКИЙ
РЕФЕРАТ ГРАФИЧЕСКИЙ
РЕФЕРАТ ИНФОРМАТИВНЫЙ
РЕФЕРАТ "ТЕЛЕГРАФНОГО СТИЛЯ"
РЕФЕРАТ УКАЗАТЕЛЬНЫЙ
РЕФЕРИРОВАНИЕ
а АННОТАЦИЯ

...

где в качестве буквенных обозначений выступают следующие:

с – термины-синонимы,

в – термины, подчиняющие заглавный термин, т.е. выше по иерархии,

н – термины, подчиненные заглавному термину, т.е. ниже по иерархии,

а – термины, ассоциированные с заглавным термином.

Особенностью тезаурусов является применяемая на практике возможность расширения словарной базы новыми ключевыми словами, появляющимися при накоплении документов в ходе эксплуатации ИПС, что порождает специфический компонент – ведение тезауруса.

Автоматизация индексирования документов

Первоначально созданные информационно-поисковые системы не предполагали автоматизации индексирования, которое осуществлялось специально подготовленными специалистами-экспертами в предметной области ИПС.

Недостатки такого подхода:

- *высокие накладные расходы* на создание и ведение документальных ИПС, т.к. требуется наличие в организационном штате высококвалифицированных специалистов-индексаторов;
- *человеческий фактор* – субъективность поисковых образов одного документа, проиндексированного разными специалистами,
- *ограничения на объем потока документов* – при большом объеме документов, возникают трудности при индексировании.

Применение системы унiterмов Тауба позволило снизить профессиональные требования к уровню квалификации индексаторов, за счет формализации процесса индексирования, но не решило всех проблем.

В рамках Кренфилдского проекта, проводившегося в конце 1950-х – начале 1960-х годов, исследовалась эффективность систем на основе УДК,

системы унитермов, фасетной системы и некоторых их разновидностей. Результаты исследований выявили:

- при повышении полноты поиска на основе использования каких-либо семантических методов при индексировании происходит снижение точности поиска и наоборот;
- небольшое отличие в показателях эффективности поиска документов с развитой семантикой индексирования и в системах на основе неконтролируемой лексики.

Последний результат привлек внимание к более простым и дешевым дескрипторным системам с неконтролируемой или слабо контролируемой лексикой, таким как унитермы и полнотекстовые системы. Сформировалось два подхода к автоматизации индексирования: индексирование в полнотекстовых системах, при котором в индекс заносится информация обо всех словах документа, и использование словаря ключевых слов, которое применяется в системах на основе информационно-поисковых тезаурусов.

Индексирование при втором подходе осуществляется путем последовательного автоматического поиска в тексте документа каждого ключевого слова. На этой основе строится и поддерживается *индекс системы*, реализующий поисковое пространство документов. Применяется два вида образования индекса – прямой и инвертированный.

Прямой тип индекса строится по схеме «Документ – термины». Поисковое пространство в этом случае представлено в виде матрицы размерностью $N \times M$ (N – количество документов, M – количество ключевых слов). Строки этой матрицы представляют поисковые образы документов.

Инвертированный тип индекса строится по обратной схеме – «Термин – документ». Поисковое пространство представлено аналогичной матрицей в транспонированной форме. Поисковыми образами документов в этом случае являются столбцы.

На основе автоматического индексирования документов по ключевым терминам также могут решаться и задачи автоматической классификации документов. Принцип решения таких задач аналогичен информационному оповещению. Для конкретного класса документов (рубрик) строится поисковый образ в виде набора терминов или их сочетаний. Поисковые образы документов из внешних источников сравниваются по определенному критерию с поисковым образом рубрики, и на основе этого принимается решение о внесении документа в базу и отнесении его содержания к предметной области ИПС» [1].

Лабораторная работа 6. Рынок электронных сделок, представленный системой заказа товара и услуг

Порядок выполнения работы

1. Используя книжный сайт, указанный в вашем варианте, осуществите поиск интересующих вас книг (не менее пяти). В отчёте отобразите их характеристики.

2. Используя сайт, предложенный в вашем варианте (соответствует порядковому номеру в журнале), найти и ознакомиться с правилами заказа и оплаты книг.

3. Осуществите заказ трёх книг (по варианту). В отчёт включите информацию о заказе (покупательской корзине) и письмо на подтверждение заказа из вашего почтового ящика.

4. Найдите в Интернете принципы классификации книг (УДК, ББК, ISBN).

Таблица 6.1 – Задание к лабораторной работе № 6.

№ вар.	Сайт	Тема
1	www.alib.ru	Эволюция
2	www.urss.ru	Кибернетика
3	www.rg.ru	Охрана труда
4	www.rsci.ru	Искусственный интеллект
5	www.lib-online.ru	Экспертные системы
6	www.bookland.ru	Взятие Берлина
7	www.ast.ru	Античная поэзия
8	www.top-kniga.ru	Экстремальное программирование
9	www.alib.ru	Фракталы
10	www.urss.ru	Астрофизика
11	www.rg.ru	Товарный знак
12	www.rsci.ru	Моделирование
13	www.lib-online.ru	Распознавание образов
14	www.bookland.ru	Философия науки
15	www.ast.ru	Геополитика
16	www.top-kniga.ru	Славянские мифы
17	www.alib.ru	Произведения Данте
18	www.urss.ru	Дифференциальная геометрия
19	www.rg.ru	Интеллектуальная собственность
20	www.rsci.ru	Нанотехнологии
21	www.lib-online.ru	Железобетонные конструкции
22	www.bookland.ru	Произведения Стругацких
23	www.ast.ru	Web-дизайн
24	www.top-kniga.ru	Настройка сервера Linux

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Тяжев, А. Т. Курс лекций по дисциплине “Информационно-поисковые системы” для студентов заочного отделения специальности 0608 [Текст] / А. Т. Тяжев. – СПб.: СПбГАСЭ, 2003.
2. Севостьянов, И. О. Поисковая оптимизация. Практическое руководство по продвижению сайта в Интернете [Текст] / И. О. Севостьянов. – 3-е изд. – СПб.: Питер, 2016.

Учебное издание

Антонюк Петр Евгеньевич

Информационно-поисковые системы

Редактор и корректор М. Д. Баранова
Техн. редактор Д. А. Романова

Учебное электронное издание сетевого распространения

Системные требования:
электронное устройство с программным обеспечением
для воспроизведения файлов формата PDF

Режим доступа: http://publish.sutd.ru/tp_get_file.php?id=202016, по паролю.
- Загл. с экрана.

Дата подписания к использованию .15. 03.2022 г. Изд. № 5263/21

Высшая школа технологии и энергетики СПбГУПТД
198095, СПб., ул. Ивана Черных, 4.