МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
ВЯТСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Биологический факультет
Кафедра биотехнологии

Допущено
к использованию в учебном
процессе по направлению
240700.62 Биотехнология
(протокол методсовета БФ
№1 от 29.08.2014)
Декан БФ_____
Мартинсон Е.А.
«29» августа 2014 г.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРАКТИЧЕСКИМ ЗАНЯТИЯМ

по дисциплине «Мировые информационные ресурсы»

г.Киров
2014 г.

Содержание

Введение

В части 2 методических указаний описываются основные модели и алгоритмы поиска информации в информационно-поисковых системах (ИПС), а также один из традиционных методов анализа документов – классификация. Перечисляются главные критерии эффективности ИПС и способы их оценки. Приводятся архитектура и состав современных ИПС, работающих в сети Интернет.

1. Модели поиска информации

Главная цель ИПС – наилучшим образом удовлетворить потребности пользователей в необходимой информации. Для реализации этой глобальной цели необходимо проделать ряд подготовительных операций, которые были подробно рассмотрены в первой части методических указаний: проанализировать информационный массив и представить его в форме, удобной для хранения и обработки. Не менее важной частью поискового аппарата ИПС является модель поиска информации. Она описывает способ и критерии сравнения запросов и документов, а также форму представления результатов этого сравнения.

Любая модель поиска тесно связана с информационно-поисковым языком. Информационно-поисковый язык (ИПЯ) – это специальный язык для формирования запросов к ИПС. Необходимость создания ИПЯ вызвана трудностями интерпретации естественного языка в компьютерной системе. Однако синтаксис информационно-поисковых языков обычно довольно прост и внешне они часто похожи на естественные. Перед использованием запросов на ИПЯ проводятся лексическая (например, удаление из запроса терминов, присутствующих в стоп-словаре), морфологическая (нормализация терминов запроса[1]), реже синтаксическая и семантическая обработки (в основном в экспериментальных системах) [].

Рассмотрим основные модели поиска информации, применяемые в ИПС.

1.1. Булева модель поиска

Наиболее распространенной моделью поиска является булева модель, позволяющая составлять логические выражения из набора терминов. Найденные документы определяются в результате описанных запросом логических операций над множеством поисковых образов документов. Пользователь получает только те доку-

[1] В базе данных ИПС термины обычно хранятся в так называемой нормальной форме. Например, для существительных - это именительный падеж единственного числа. Одновременная нормализация терминов запросов и документов позволяет существенно упростить процесс их сравнения при поиске.

менты, чьи наборы терминов точно совпадают с соответствующими комбинациями терминов запроса.

Поисковые образы запросов связывают термины с помощью булевых операторов ("И" – "AND", "ИЛИ" – "OR", "И НЕ" – "AND NOT"). Эти операции производятся над множествами документов, содержащих тот или иной термин, определенный запросом. Для обозначения объединения множеств ("ИЛИ" в запросе) применяется символ □, пересечения множеств ("И" в запросе) – □, разности множеств ("И НЕ" в запросе) – \.

Например, оператор "И", соединяющий два термина запроса, означает следующее. Из множества всех документов нужно сначала выбрать два подмножества. Одно из них содержит первый термин запроса, а другое – второй. Затем определяется общая часть (пересечение) этих подмножеств, то есть те документы, в состав которых одновременно входят и первый, и второй термины из запроса.

Рассмотрим, например, такой запрос:

(((Microsoft *and* Word) *or* (Microsoft *and* Excel))
***and* Macintosh) *and not* Windows**

В данном случае выражение на ИПЯ означает следующее: нужно найти все документы, которые одновременно содержат либо сочетание "Microsoft Word", либо сочетание "Microsoft Excel", а также содержат слово "Macintosh", но не содержат слово "Windows".

Этот запрос можно разбить на две части:

1. **Microsoft *and* Word *and* Macintosh *and not* Windows**
2. **Microsoft *and* Excel *and* Macintosh *and not* Windows**

Выполнение первого запроса происходит в два этапа. Сначала находятся все документы, содержащие термины "Microsoft", "Word" и "Macintosh". Затем из найденных документов отсеиваются те, которые содержат слово "Windows". Второй запрос выполняется аналогично. В конце производится объединение результатов работы первой и второй частей исходного запроса.

Часто пользователь строит свой запрос, не используя каких-либо логических операторов, и просто перечисляет ключевые слова. В таком случае обычно предполагается, что все термины соединены логической операцией "И".

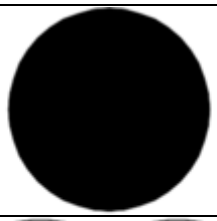
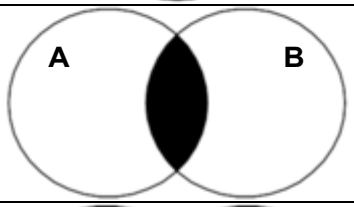
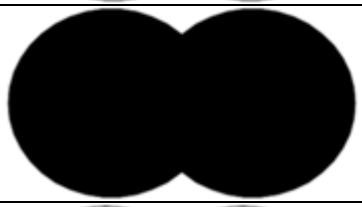
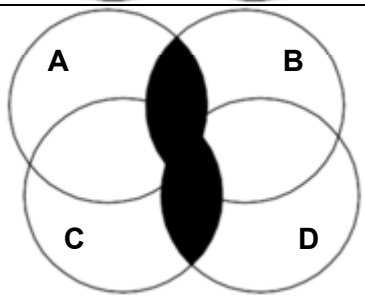
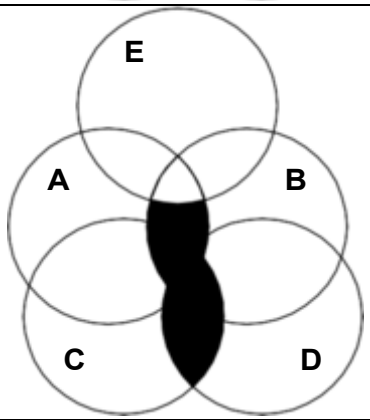
В некоторых поисковых системах вместо булевых операторов язык запросов позволяет использовать различные знаки. Так, знак "+" эквивалентен оператору "И", знак "-" – оператору "И-НЕ" и т. д.

В процессе поиска из исходного информационного массива выделяется часть, которая содержит найденные документы, соответствующие комбинациям терминов

запроса. Какого-либо упорядочения (например, ранжирования по релевантности) не проводится: все выданные документы считаются одинаково важными.

Несколько типичных булевых стратегий поиска изображено в табл.1.

Таблица 1. Поиск с использованием булевых операторов

Формулировка запроса. Термины	Операции с множествами документов	Результат поиска
(a)	A	
$(a \ AND \ b)$	$A \cap B$	
$(a \ OR \ b)$	$A \cup B$	
$(a \ AND \ b)$ $OR \ (c \ AND \ d)$	$(A \cap B) \cup (C \cap D)$	
$((a \ AND \ b) \ OR$ $(c \ AND \ d)) \ AND \ NOT \ e$	$((A \cap B) \cup (C \cap D)) \setminus E$	

Здесь a, b, c, d, e – термины, из которых состоят запросы, а A, B, C, D, E – множества документов, содержащих эти термины (например, A - это множество документов, содержащих термин a, и т. д.).

ИПС, работающие с булевой моделью поиска, имеют ряд недостатков [,].

1.Обычные булевы запросы затрудняют варьирование глубины поиска с целью выдачи большего или меньшего количества документов в зависимости от требований пользователя. Для получения желаемого уровня эффективности необходимо найти правильную формулировку запроса: не слишком широкую и не слишком узкую. Оператор AND может привести к резкому сокращению числа найденных документов, а оператор OR, напротив, может чрезмерно расширить запрос и выделить нужную информацию из информационного шума будет трудно. Результат поиска также сильно зависит от того, насколько типичными для базы данных ключевых слов являются термины запроса. Поэтому для успешного применения булевой модели следует хорошо ориентироваться в предметной лексике. Для повышения результативности создаются специальные словари - тезаурусы, которые содержат информацию о связи терминов друг с другом.

2.При использовании булевой логики нельзя получить эффект от функций совпадения векторов, которые дают непрерывный спектр совпадений (полных, частичных или нулевых) между запросами поисковыми образами документов. Это обстоятельство приводит к жесткому требованию "все или ничего" на выходе.

3.Еще одним минусом является тот факт, что множество выданных документов не может быть представлено пользователю в ранжированном[1] виде, например в порядке уменьшения сходства между документом и запросом. Документ либо полностью соответствует запросу, либо не соответствует совсем. Эта проблема может быть решена с помощью взвешенного булева поиска, при котором производится частичное ранжирование с использованием весов терминов W_i. Результаты поиска располагаются в порядке уменьшения весов совпавших терминов [, ,].

Несмотря на описанные недостатки, булева модель поиска широко применяется в современных ИПС из-за простоты ее реализации.

1.2. Функции подобия "документ-запрос"

Негативные свойства, характерные для булевого алгоритма поиска, обусловлены употреблением в запросе логических операторов, приводящих к жестким условиям поиска. Одним из решений этой проблемы является отказ от их использования и, как следствие, разработка каких-либо других алгоритмов поиска.

Многие современные ИПС реализуют модели поиска информации, основанные на вычислении мер близости документов и запросов []. ИПЯ, используемые в таких

[1] Ранжирование – упорядочение результатов поиска по некоторому критерию соответствия их информационной потребности пользователя.

моделях, называются языками типа "найти похожее" (языки типа "Like This"). В этих языках необязательно формулировать запросы с помощью булевых операторов.

Для вычисления меры подобия документов и запросов существует более тридцати различных алгоритмов [,]. На сегодняшний день используется лишь несколько из них. Мы рассмотрим четыре алгоритма []:

–расширенного булевого поиска,

–наибольшего цитирования,

– $TF \times IDF$ алгоритм,

–расширенный векторный алгоритм поиска.

Алгоритмы расширенного булевого поиска и наибольшего цитирования основаны на метаинформации гипертекстовых страниц. $TF \times IDF$ алгоритм использует статистические частотные оценки встречаемости терминов. Расширенный векторный алгоритм работает как с частотными оценками, так и с гипертекстами.

Прежде всего введем некоторые обозначения:

M – число терминов в запросе.

q – запрос, состоящий из M терминов (вектор запроса).

Q_j – j -й термин запроса, $j = \overline{1, M}$.

N – число документов в информационном массиве.

P_i – i -й документ (поисковый образ i -го документа), $i = \overline{1, N}$.

$R_{i,q}$ – релевантность (мера близости) P_i по отношению к запросу q.

$C_{i,j}$ – величина, характеризующая наличие Q_j в P_i, определяемая по формуле

$$C_{i,j} = \begin{cases} 0, Q_j \notin P_i \\ 1, Q_j \in P_i \end{cases}. \qquad (1.1)$$

Для повышения качества поиска в выражении (1.1) вместо единицы можно также использовать вес термина в документе $W_{i,j}$.

$I^L_{i,k}$ – величина, характеризующая наличие гиперссылки из P_k в P_i (входя-щей[1]), $I_{L^{i,k}} = 1$, если она есть.

[1] IL – англ. Incoming Hyperlink – входящая гиперссылка.

$OL_{i,k}$ – величина, характеризующая наличие гиперссылки из P_i в P_k (исходящей ссылки), $OL_{i,k} = 1$, если она есть.

1.2.1. Алгоритм расширенного булевого поиска

Алгоритм расширенного булевого поиска основан на булевой модели, причем расширением является возможность ранжировать найденные документы по числу терминов запроса, которые в них встречаются. Такую модель поиска можно рассматривать как упрощенную модель поиска в нечетких множествах [] в противоположность строгим множествам булевого поиска.

Релевантность документа P_i по отношению к запросу q рассчитывается как

$$R_{i,q} = \sum_{j=1}^{M} C_{i,j}. \qquad (1.2)$$

Алгоритм расширенного булевого поиска использует модель (1.2) не только для данного документа, но и для соседних с ним, учитывая частоту появления в них слов запроса. Такое становится возможным в среде гипертекстовых документов. Предполагается, что если два документа связаны гиперссылкой, то между ними должна существовать и некоторая семантическая (смысловая) связь.

Практически это выглядит следующим образом. Если документ P_i не содержит термина запроса Q_j, но связан с другими документом P_k, в который этот термин входит, то полагают, что документ P_i содержит термин Q_j. Однако при этом во время ранжирования документу P_i приписывается меньший вес, чем если бы он на самом деле содержал термин Q_j. Алгоритм определения релевантности документа P_i и запроса q принимает вид

$$R_{i,q} = \sum_{j=1}^{M} I_{i,j},$$

где $I_{i,j}$ определяется следующим образом:

$$I_{i,j} = \begin{cases} c, \text{ если } C = 1, \\ c_2, \text{ если существует такое } k, \text{ что } C_{k,j} = 1 \text{ и } (IL_{i,k} + OL_{i,k}) > 0, \\ 0, \text{ во всех других случаях.} \end{cases}$$

Здесь c_1 и c_2 – положительные константы, причем $c_1 > c_2$.

1.2.2. Алгоритм наибольшего цитирования

Этот алгоритм также использует информацию о гиперссылках между документами. Мера релевантности каждой страницы P_i определяется суммой числа терминов запроса, содержащихся на других страницах, которые имеют ссылку на данную:

$$R_{i,q} = \sum_{k_i} \left(C_{i,j} \left(IL_{i,k} \sum_{j=1}^{M} C_{k,j} \right) \right).$$

Цель данного алгоритма – приписать большие веса тем документам в множестве найденных, которые цитируются (на которые ссылаются другие документы) чаще всего. Аналогичный подход применяется также в ряде других алгоритмов, в частности, в алгоритме PageRank, который используется в информационно-поисковой системе Интернет Google [].

1.2.3. Векторный алгоритм поиска

Векторный алгоритм поиска, называемый $TF \times IDF$-алгоритмом, является одним из самых распространенных. Он основан на векторной модели информационного массива, в которой для определения меры близости документов и запросов используется значение косинуса угла между их векторами в многомерном пространстве информационного массива [,].

Запросы и документы в векторной модели представляют множествами наборов взвешенных терминов. Вектор запроса q в таком случае будет выглядеть так:

$$q = (W_1, W_2, \ldots, W_j, \ldots, W_M),$$

где W_j – вес j-ого термина в запросе (вес термина Q_j в запросе q).

Вектор документа P_i можно представить как

$$P_i = (W_{1,i}, W_{2,i}, \ldots, W_{j,i}, \ldots, W_{M,i}),$$

где $W_{j,i}$ – вес термина Q_j в документе P_i.

Функция совпадения векторов запроса и документа имеет вид:

$$R_{i,q} = \frac{\sum_{j=1}^{M} \left(W_j W_{j,i} \right)}{\sqrt{\sum_{j=1}^{M} \left(W_j \right)^2 \sum_{j=1}^{M} \left(W_{j,i} \right)^2}}. \qquad (1.3)$$

Числитель дроби (1.3) определяет скалярное произведение векторов документа и запроса, знаменатель – произведение их длин, а релевантность $R_{i,q}$ – косинус угла между этими векторами в Евклидовом пространстве. Весовые коэффициенты терминов запроса будут постоянными от документа к документу. Поскольку для оценки релевантности обычно важно знать изменение меры подобия документов, а не ее абсолютное значение, а также для ускорения процесса вычислений, характеристики запроса в выражении (1.3) можно не учитывать:

$$R_{i,q} = \frac{\sum_{j=1}^{M} W_{j,i}}{\sqrt{\sum_{j=1}^{M} \left(W_{j,i} \right)^2}}. \qquad (1.4)$$

Вес терминов $W_{j,i}$ в выражении (1.4) обычно вычисляется по формулам, приведенным в части 1 методических указаний. В частности, окончательное выражение для релевантности q и P_i, описывающее $TF \times IDF$ -алгоритм, может иметь вид

$$R_{i,q} = \frac{\sum_{j=1}^{M} \left(0.5 + 0.5 \frac{(TF)_{i,j}}{(TF)_{i,\max}} \right) (IDF)_j}{\sqrt{\sum_{j=1}^{M} \left(0.5 + 0.5 \frac{(TF)_{i,j}}{(TF)_{i,\max}} \right)^2 \left((IDF)_j \right)^2}}, \qquad (1.5)$$

где $(TF)_{i,j}$ – частота термина Q_j в документе P_i;

$(TF)_{i,\max}$ – частота максимально часто встречающегося термина в P_i;

$(IDF)_j$ – обратная документная частота, вычисляемая по формуле

$$(IDF)_j = \log \left(\frac{N}{N_{i,j}} \right).$$

Вычисление длины вектора документа (знаменатель выражения (1.5)) занимает очень много времени. Поэтому часто применяют упрощенный $TF \times IDF$-алгоритм:

$$R_{i,q} = \sum_{j=1}^{M} \left(0.5 + 0.5 \frac{(TF)_{i,j}}{(TF)_{i,\max}} \right) (IDF)_j . \qquad (1.6)$$

Практика показывает, что упрощенный алгоритм (1.6) при поиске в Интернете является более эффективным, чем полный алгоритм (1.5).

1.2.4. Расширенный векторный алгоритм поиска

Этот алгоритм является комбинацией векторного алгоритма и алгоритма наибольшего цитирования. Сначала релевантность каждого документа вычисляется по $TF \times IDF$ -алгоритму, а затем корректируется с учетом связанных документов.

Мера близости документа P_i и запроса q рассчитывается по формуле

$$R_{i,q} = S_{i,q} + \sum_{j=}^{N} \left(\alpha \cdot IL_{i,j} \cdot S_{j,q} \right) .$$

где $S_{i,q}$ и $S_{j,q}$ – релевантность документов, полученная по формуле (1.6);

α – постоянный весовой коэффициент $(0 < \alpha < 1)$.

2. Классификация документов

Во время поиска часто бывает важно получить по возможности наибольшее значение полноты, то есть выдать максимальную часть релевантных документов, имеющихся в массиве. Исчерпывающий поиск может понадобиться, например, экспертам организации, регистрирующей изобретения, которым необходимо составить обзор всех существующих патентов. Увеличение числа релевантных документов обычно приводит к выдаче дополнительных нерелевантных документов, то есть снижается его точность (см. часть 1 методических указаний).

Для улучшения полноты поиска необходимы дополнительные совпадения терминов запроса и документа. Это достигается использованием дополнительных терминов-заместителей []. Термины-заместители либо добавляются к уже существующим терминам запросов и документов, либо используются вместо них. Наиболее известным методом здесь является применение словаря синонимов (тезауруса), в котором термины сгруппированы в классы синонимии (классы эквивалентности).

С помощью тезауруса можно заменить каждый имеющийся в начальный момент поиска термин идентификаторами соответствующих классов тезауруса. При использовании другого подхода идентификаторы этих классов можно добавлять к исходным терминам. В любом случае цель состоит в том, чтобы получить дополни-

тельные совпадения для тех терминов запроса и документа, которые отнесены к одним и тем же классам тезауруса. Сами эти термины могут быть и различными [].

В ИПС в основном применяется два типа классификаций []: терминов и документов.

Целью классификации терминов является группировка терминов в синонимические классы в расчете повысить вероятность совпадения терминов запроса и документа. Классификация документов способна улучшить результаты и оперативность поиска за счет обращения только к определенным частям информационного массива. Эти два типа классификаций взаимосвязаны: присваиваемые документам термины при формировании их поисковых образов служат основой для построения классов, получаемых в результате группировки документов.

При хорошей классификации терминов обычно удается сгруппировать различные низкочастотные родственные термины в общие классы тезауруса. Термины, входящие в один класс, могут заменять друг друга в процессе поиска, следовательно, можно ожидать улучшения полноты выдачи. Классификации документов позволяют сузить область поиска до наиболее существенных классов документов и обеспечить высокую точность. При совместном использовании систематизированных массивов данных и тщательно проработанного тезауруса можно получить высокие показатели и по полноте, и по точности поиска.

В основе любой классификации лежит принцип распределения информационных объектов (терминов или документов) по некоторым классам. Совокупность таких классов называется классификатором, а сами классы – разделами классификатора, или рубриками. Классификаторы обычно разрабатываются вручную []. Примерами классификаций могут служить общепринятые библиотечные классификации УДК (универсальная десятичная классификация) и ББК (библиотечно-библиографическая классификация) [].

Класс определяется как множество терминов, обозначающих некоторую предметную область. В процессе классификации каждому информационному объекту для обозначения его смыслового содержания (тематики) приписывается идентификатор какого-либо класса [].

Разбиение на предметные классы или рубрики должно быть предсказуемым, а подчиненные тематические классы легко отличимы от вышестоящих. От четкости такой иерархической структуры зависит эффективность регулирования глубины поиска путем расширения или сужения запроса.

Маловероятно, чтобы можно было найти такую структуру, которая могла бы удовлетворять этим требованиям. Строго заданные иерархические отношения меж-

ду тематическими классами призваны подчеркнуть определенные типы предметных ассоциаций и одновременно пренебречь другими. Статичный характер общепринятых классификационных схем порождает проблемы в случае расширение предметных областей и развития знаний. Существующие иерархические схемы весьма сложны, и на практике часто оказываются обязательными ручные (неавтоматические) процессы классификации. Это приводит к тому, что согласованности между разными системами классификации и поиска в процессах анализа содержания и распределения документов по рубрикам добиться трудно [, ,].

2.1. Основные свойства классификации

В ИПС процесс классификации документов происходит во время их индексирования. Термины запроса распределяются по рубрикам классификатора непосредственно во время поиска. В обоих случаях документы и термины составляют множество классифицируемых объектов. Если множество объектов необходимо сопоставить множеству классов, обычно требуется, чтобы получающаяся при этом классификация обладала следующими свойствами []:

1.Классификация должна быть *корректно определенной* так, чтобы для любого заданного множества данных получался один результат.

2.Результаты классификации не должны зависеть от порядка обработки объектов (*независимость от порядка*), то есть любая перестановка анализируемых объектов не должна влиять на результат классификации.

3.Классификация должна быть *устойчивой*: незначительные изменения данных должны вызывать незначительные изменения результатов классификации.

4.Классификация должна быть *независимой от масштаба*, поскольку умножение на константу значений характеристик, идентифицирующих объекты (идентификаторов классов), не должно влиять на классификацию.

5.Объекты, обладающие большим сходством, не должны оказываться отнесенными к разным классам.

Первые два свойства (корректность определения и независимость от порядка) взаимосвязаны. Они могут быть обеспечены только при условии предварительного анализа всех возможных подмножеств объектов, удовлетворяющих классификационным критериям. Однако при большом количестве объектов, подлежащих классифицированию, такой исчерпывающий анализ может потребовать значительных затрат времени, что имеет место, например, в сети Интернет.

Если первый и второй критерии не удовлетворяются, то особую важность приобретает критерий устойчивости классификации. Он гарантирует, что добавление

новых свойств объектов, устранение уже выделенных свойств, а также исправление незначительных ошибок вызовут лишь незначительные изменения в самих классах.

В классификациях, используемых в ИПС, обычно стараются получать устойчивые классы терминов и документов особенно потому, что векторы свойств, характеризующие объекты, не всегда точны и надежны. Это связано, например, с тем, что некоторые термины, несущие важную смысловую нагрузку, могут игнорироваться при автоматическом анализе содержания документов.

Системы классификации имеют также ряд формальных свойств []. Если все члены одного и того же класса обладают одним общим признаком, то классификация называется монотетической. В противном случае классификация становится политетической. Классы могут быть непересекающимися, где объекты относятся самое большее к одному классу, и пересекающимися. Наконец, классификация может быть упорядоченной путем установления систематических отношений между различными классами, а может быть и неупорядоченной.

В процессе разработки и проектирования систем классификации во всех случаях предпочтительнее менее жесткие требования. Обычно ни документы, ни термины не бывают определены настолько точно, чтобы имело смысл строить монотетические классификации терминов или документов. По этой же причине наилучшими классами должны считаться пересекающиеся классы, чтобы элемент (термин или документ) мог включаться более чем в один класс.

В некоторых случаях целесообразно создание либо упорядоченных классификаций терминов (иерархий терминов), либо упорядоченных классов документов. Однако, когда не налагается никаких специальных требований, неупорядоченная классификация, как правило, дает более адекватное деление на классы. Таким образом, в общем случае наиболее предпочтительными являются политетические пересекающиеся неупорядоченные классификации.

В любой ИПС существует тесная взаимосвязь между индексированием и классификацией. Часто два этих процесса осуществляются параллельно. Целью классификации терминов является формирование для каждого термина дополнительных заместителей. Эти же термины используются и для идентификации документов.

Представление и классификация документов в ИПС также связаны между собой. При индексации каждому документу обычно сопоставляется некоторый набор индексационных терминов. Поэтому фактически используемые термины непосредственно оказывают влияние как на классификацию терминов, так и на классификацию документов. Например, во время автоматической классификации документов определяется мера близости между классифицируемым документом и некоторым

эталонным документом, который заведомо принадлежит какому-либо определенному классу. Эта мера часто вычисляется в зависимости от терминов, входящих в векторы этих документов, например по формуле (1.3). Поэтому классы документов непосредственно зависят от методов индексирования [,].

2.2. Формирование рубрик

Типичный процесс формирования рубрик (классов) включает три основных процесса (рис. Рис. 1) [].

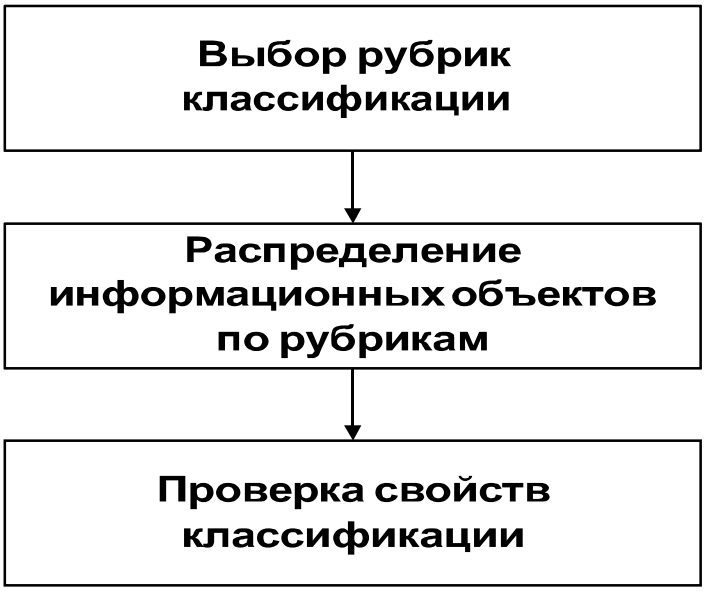


Рис. 1. Процесс формирования рубрик

Во время начального процесса происходит определение рубрик. Часто эта операция сводится к выбору (в качестве центра исходных классов) объектов, размещенных в плотных зонах пространства информационных объектов. Такими зонами обычно считаются те, в окрестностях которых имеется большое количество подобных объектов.

В процессе распределения информационные объекты систематизируются и распределяются по имеющимся рубрикам путем отнесения объектов к тем классам, с которыми они имеют достаточно высокий коэффициент подобия.

Завершающий этап связан с выполнением условий, при которых данный класс считается окончательным и полным. Здесь устанавливается, удовлетворяют ли сформированные рубрики заданному критерию классификации (например, обладают ли они описанными в предыдущем параграфе свойствами).

Существует два основных метода классификации []:

1.Порождающие методы классификации по принципу снизу вверх.

2.Методы разбиения по принципу сверху вниз.

При использовании порождающих методов все объекты первоначально считаются несгруппированными. Формирование групп выполняется снизу вверх путем объединения объектов.

Методы разбиения по принципу сверху вниз подразумевают, что все объекты первоначально относятся к одному глобальному классу. Затем этот класс разбивается на более мелкие подклассы, которые в свою очередь могут делиться на еще более мелкие подклассы вплоть до образования окончательных классов.

В действующих системах также используется смешанный метод классифицирования по принципу сверху вниз. Количество исходных классов в таком случае задается заранее, и первоначальное деление объектов корректируется путем перегруппировки объектов. Целью перегруппировки является повышение качества рубрик таким образом, чтобы связанность классов стала максимальной, а подобие объектов, относящихся к разным группам, – минимальным.

Большая часть методов классификации по принципу сверху вниз устроена таким образом, что они могут использоваться и для образования иерархических структур классов. При поуровневом построении классификации формируются классы, являющиеся подмножествами или компонентами какого-либо класса более высокого уровня. В результате образуется структура в виде дерева. Корень такого дерева (верхний уровень) содержит глобальный класс высшего уровня, представляющий все информационное пространство. Листья (нижний уровень) соответствуют конечным рубрикам документов или группам терминов.

При некоторых методах классификации по принципу снизу вверх также формируются иерархические структуры. Неиерархическими структурами считаются такие структуры, в которых между сформированными классами не выполняются свойства формального включения. При построении иерархии классов терминов в виде дерева часто стараются в нижней части помещать узкие специфичные термины, а в верхней – термины более общего характера.

На практике особенно во время ручной классификации часты случаи, когда документ или термин может быть одновременно отнесен к нескольким классам. В таких ситуациях используются различные перекрестные ссылки [].

Информация о документах данной тематической направленности помещается в некоторый базовый раздел, а остальные классы, к которым также можно было бы отнести эти документы, содержат соответствующие ссылки. В описание пересекающихся классов добавляют ссылку типа "смотри", которая направляет пользователя к рубрике, признанной специалистами по классификации базовой.

Например, информация о картах стран может быть размещена в разделах "Наука–География–Страна", "Экономика–География–Страна" или "Справочники–Карты–Страна". Специалисты по классификации принимают решение о том, что сведения о картах стран размещаются в рубрике "Экономика–География–Страна". Тогда в остальные два раздела добавляется ссылка на данный.

Если выбор базового класса вызывает у специалистов по классификации затруднения, то вероятность отнесения объекта к тому или иному похожему (синонимическому) классу практически одинакова. В этих случаях применяются ссылки типа "смотри также". Они направляют пользователей системы к разделам, которые, возможно, содержат описания интересующих их документов.

3. Эффективность поисковых систем

3.1. Критерии эффективности

Эффективность любой информационной системы определяется ее способностью служить тем целям, для которых она была разработана. Поскольку ИПС существует в конечном счете для удовлетворения информационных потребностей, критерии ее эффективности определяются пользователями [,].

Существует два направления оценки качества работы поисковых систем. В одном случае анализируется отдельно взятая ИПС, в другом – определяются характеристики эффективности по сравнению с другими системами. Оценивать эффективность ИПС можно либо количественно, либо качественно.

При первом типе оценки качества выводы должны быть тщательно проверены и подтверждены экспериментальными доказательствами, а рассматриваемая поисковая система должна быть подвергнута комплексным испытаниям. Программа испытаний при этом должна учитывать большинство параметров и переменных системы и основываться на убедительном теоретическом базисе. Испытания второго типа не обязательно приводят к бесспорно доказуемым результатам. Подобные испытания часто можно проводить, используя имитационные методы. Практика показывает, что многое о качестве работы системы удается узнать из серии качественных экспериментов, даже если отсутствует полная уверенность в применимости полученных результатов к конкретным эксплуатационным условиям [].

Оценка поисковых систем может производиться на нескольких уровнях [,]:

–инженерный уровень исследует характеристики эффективности программного и аппаратного обеспечения: надежность, гибкость, скорость вычислений, а также эффективность применяемых поисковых алгоритмов;

–на уровне входа изучаются вопросы, связанные с входной информацией и внутренним содержимым системы, в частности, о степени полноты имеющихся информационных ресурсов в определенной области;

–уровень обработки рассматривает вопросы качества работы алгоритмов поиска, обоснованности применяемых методов и подходов;

–на уровне выхода исследуется взаимодействие пользователя с системой и работа с полученными результатами: вид представления найденных документов, оценка механизмов обратной связи и т. д;

–уровень применимости системы анализирует возможности использования результатов поиска для решения стоящей перед пользователем задачи и степень полезности этих результатов;

–социальный уровень исследует влияние системы на ее окружение, а именно на эффективность принятия решений, производительность труда и т. д.

В зависимости от целей и условий оценки эффективности можно выбрать множество методов исследования. На практике часто применяется метод макрооценки. Анализируемая ИПС рассматривается в таком случае как черный ящик, то есть ее структура не принимается во внимание, а акцент делается на затраты времени и ресурсов на уровне входа и получение нужных документов на уровне выхода [].

Необходимо отметить, что испытание ИПС в любом случае должно производиться с использованием набора запросов, отражающего реальные типы запросов, которые в действительности поступают в условиях эксплуатации системы. Одновременно оценка релевантности найденных разными системами документов должна проводиться одними и теми же пользователями (экспертами) [,].

Принято выделять несколько основных критериев эффективности ИПС:

1.Полнота поиска - способность ИПС выдавать все релевантные документы.

2.Точность поиска - способность ИПС отсеивать нерелевантные документы.

3.Усилия, затрачиваемые на формулирование запросов, взаимодействие с системой и просмотр выдаваемой информации.

4.Форма представления найденной информации.

5.Полнота информационного массива, то есть степень охвата всех релевантных информационных ресурсов, интересующих пользователей.

Некоторые из этих критериев можно измерить довольно легко. Например, затраты труда пользователей можно выразить через время, необходимое для формулирования запроса, диалога с системой и просмотра полученной информации. Так же непосредственно можно оценить форму представления документов. Определение полноты охвата информационного массива может вызывать затруднения, если

заранее неизвестно количество документов, представляющих интерес в данной предметной области. Это особенно характерно для глобальных ИПС сети Интернет. Наиболее трудным как принципиально, так и практически, является определение мер полноты и точности, то есть оценка качества результатов поиска.

3.2. Полнота и точность поиска

Коэффициент полноты – это доля полученных релевантных документов по сравнению с их общим количеством в поисковом массиве. Коэффициент точности – это доля релевантных документов среди выданных.

Введем обозначения []:

a – количество полученных в результате поиска релевантных документов,

b – количество нерелевантных документов, выданных ИПС,

c – число релевантных документов в поисковом массиве, не выданных ИПС,

d – число невыданных релевантных документов.

Табл. 2 иллюстрирует подобное разделение документов на подмножества.

Таблица 2. Разделение документов в процессе поиска

Документы	Релевантные	Нерелевантные	Всего
Выданные	a	b	$a+b$
Невыданные	c	d	$c+d$
Всего	$a+c$	$b+d$	$a+c+b+d$

Тогда коэффициент полноты[1] R и коэффициент точности[2] P можно определить по формулам:

$$R = \frac{a}{a+c}, \tag{3.7}$$

$$P = \frac{a}{a+b}. \tag{3.8}$$

Часто используются дополнительные меры оценки – коэффициент выпадения[3] F, характеризующий количество возвращаемых системой нерелевантных документов, и коэффициент ошибки[4] E, описывающий правильность определения поисковой системой релевантности документов:

[1] Полнота – англ. Recall.
[2] Точность – англ. Precision.
[3] Выпадение – англ. Fallout.
[4] Ошибка – англ. Error.

$$F = \frac{b}{b+d},$$

$$E = \frac{b+c}{a+b+c+d}.$$

Если исследовать эффективность поисковой системы с помощью нескольких запросов (обозначим общее число запросов через k), то для данного запроса i коэффициенты полноты R_i и точности P_i можно записать в виде:

$$R_i = \frac{a_i}{a_i + c_i}, \quad (3.9)$$

$$P_i = \frac{a_i}{a_i + b_i}. \quad (3.10)$$

Из уравнений (3.9) и (3.10) можно получить среднюю величину, которая отражает эффективность системы, ожидаемую для случая среднего пользователя. Для этого возьмем среднее арифметическое по k выборочным запросам:

$$R_{RL} = \frac{1}{k} \sum_{i=1}^{k} \frac{a_i}{a_i + c_i},$$

$$P_{RL} = \frac{1}{k} \sum_{i=1}^{k} \frac{a_i}{a_i + b_i}.$$

Поскольку значения коэффициентов полноты R_i и точности P_i определяются однозначно для каждого из запросов пользователей, это позволяет вычислить средние значения для фиксированных интервалов полноты. Кривая, полученная в результате усреднения, называется кривая "полнота-точность" поисковой системы (рис. Рис. 2). Левый край этой кривой соответствует узким, специфичным формулировкам запросов, а правый - определяется широкими, общим запросами.

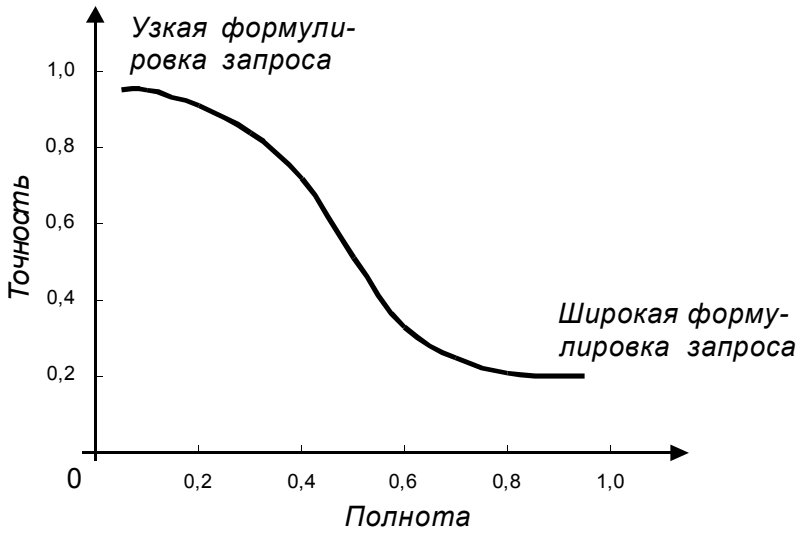


Рис. 2. Кривая "полнота-точность"

Кривые “полнота-точность” могут использоваться для оценки качества работы либо нескольких ИПС, либо одной, работающей в разных условиях. При этом кривые, полученные для двух систем, могут быть наложены на один график, что позволяет определить, какая из систем лучше и в какой степени []. Очевидно, что кривая, расположенная ближе к правому верхнему углу графика (рис. Рис. 2), где полнота и точность максимальны, указывает на лучшее качество работы.

В идеальной ИПС коэффициенты полноты и точности равны единице. В реальных поисковых системах коэффициент полноты поиска может достигать значений 0,7 – 0,9, а коэффициент точности находится в интервале 0,1 – 1,0 [].

В дополнение к стандартным мерам полноты (3.7) и точности (3.8), значения которых зависят от размера множества выданных документов, можно использовать показатели, не зависящие от выданного множества. В частности, для систем, в которых полученные документы ранжируются в порядке уменьшения сходства между документом и запросом, существуют меры оценки, основанные на рангах релевантных документов. Такие функции, называемые нормализованной полнотой и нормализованной точностью, имеют вид:

$$R_{норм} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=}^{n} i}{n(N-n)}$$

$$P_{норм} = 1 - \frac{\sum_{i=1}^{n} \log r_i - \sum_{i=1}^{n} \log i}{\log \frac{N!}{n!(N-n)!}}.$$

где n – количество релевантных документов в массиве; N - объем всего массива документов; r_i – ранг i -го релевантного документа в случае, когда документы расположены в порядке уменьшения их сходства с запросом [,].

В идеальной системе все релевантные документы находятся в верхней части списка выданных документов, то есть $r_i = i$ при $1 \le i \le n$. Нормализованные полнота и точность равны в этом случае единице.

3.3. Недостатки основных характеристик

Применение мер полноты и точности для оценки эффективности поиска имеет ряд ограничений. Во-первых, из определений (3.7) и (3.8) ясно, что измерения R и P обычно привязаны к конкретному массиву документов и конкретному множеству запросов. В пределах такой фиксированной среды имеется возможность варьировать методы и язык индексирования, методику поиска, и в результате можно определить, как эти изменения влияют на работоспособность системы с точки зрения полноты и точности. Однако абсолютно неприемлемо сравнивать показатели полноты и точности совершенно различных систем, основанных на разных массивах документов, наборах запросов и группах пользователей.

Например, полнота и точность в той или иной степени зависят от размера информационного массива и среднего количества релевантных документов, находящихся в массиве. Можно предполагать, что по мере роста объема массива полнота и точность будут ухудшаться, если только количество релевантных документов не будет увеличиваться пропорционально размеру массива. То же справедливо для случая, когда при анализе эффективности используется новое множество запросов, для которого среднее количество релевантных документов меньше, чем для первоначального множества запросов [,].

Во-вторых, коэффициенты полноты и точности несложно вычислить только в том случае, если каждый документ можно однозначно отнести либо к множеству релевантных, либо нерелевантных. Когда размер информационного массива сравнительно невелик (в локальных ИПС или тестовых наборах документов глобальных ИПС), часто имеется возможность получить однозначные оценки релевантности каждого документа по отношению к конкретным запросам.

В более крупных массивах исчерпывающие оценки релевантности обычно невозможны. Здесь для получения достоверных показателей полноты бывает необходимо оценить как общее число релевантных документов в массиве, так и позицию (ранг) релевантных документов в списке выданных. Это можно сделать

методами случайных выборок. Список релевантных документов может быть получен на основе оценок релевантности только выданного множества документов.

Кроме того, классификация релевантности на основе бинарной логики не вполне адекватна понятию релевантности. Документ может быть частично релевантен информационной потребности. Возможна ситуация, когда информационную потребность удовлетворяет совокупность из нескольких документов, и при этом релевантность каждого из них можно охарактеризовать некоторым числом. При этом использование формальной релевантности, значение которой рассчитывается для каждого документа в ходе выполнения поискового алгоритма, является неприемлемым для анализа качества работы системы с точки зрения потребителей [].

Специфика сети Интернет также накладывает существенные ограничения на применение показателей полноты и точности для оценки эффективности поиска [].

К факторам, влияющим на расчет этих характеристик, относятся очень большое количество документов, значительная доля релевантных документов, ограниченность возможностей пользователя. Остановимся на них более подробно.

В настоящее время в сети Интернет находится несколько миллиардов документов, причем их число постоянно увеличивается. В массивах поисковых образов наиболее мощных ИПС содержатся сведения о части этих документов, которая составляет по разным оценкам от трех до восьми миллиардов документов, по состоянию на конец 2002 года.

При определении коэффициента полноты поиска используется количество релевантных документов, не выданных ИПС (3.7). Как отмечалось выше, оценить это количество можно на основе изучения некоторой выборки этих документов. Однако построение такой выборки вызывает существенные затруднения из-за невозможности охвата всех документов. Недостаточная представительность выборки обуславливает появление значительной систематической погрешности при расчете числа невыданных релевантных документов.

В последние 5-8 лет происходит интенсивный процесс перевода в электронную форму и размещения в сети Интернет основного массива наиболее значимых из созданных ранее печатных документов. В тоже время многие вновь создаваемые документы практически сразу размещаются в сети. В результате большинству возникающих у пользователя информационных потребностей соответствуют десятки тысяч релевантных документов, размещенных в сети. Вместе с тем релевантная информация во многих документах совпадает, и пользователю достаточно просмотреть лишь несколько из них. Таким образом, высокое значение коэффициента полноты не является актуальным и может приближаться к нулю в случае успешного поиска. Сле-

довательно, этот коэффициент в данном случае не является адекватным описанием эффективности информационного поиска [,].

Ограниченность возможностей пользователя состоит в том, что практически всегда на просмотр и изучение результатов поиска выделяется ограниченное время. Более половины пользователей изучают только первые 10 документов, выдаваемых поисковой системой, а пятая часть – первые 20 документов. Поэтому при оценке качества поиска следует учитывать только ту часть результатов поиска, которая реально может быть изучена, а не весь список выданных системой документов [,].

Подводя итог, отметим, что в настоящее время не существует универсальной меры эффективности ИПС, которая бы устраняла описанные недостатки. Наличие большого количества характеристик, которые с трудом поддаются формализации приводит к тому, что единой теории оценки ИПС до сих пор нет, а предлагаемые методы оценки носят экспериментальный характер. Тем не менее оценка качества поиска является одним из основных факторов, влияющих на развитие ИПС [, ,].

4. Современные информационно-поисковые системы

Разнообразные технологии и методы, созданные за годы развития теории и практики информационного поиска, находят свое применение в современных ИПС.

Наряду с классическими библиотечными ИПС, которые продолжают совершенствоваться, интенсивное развитие происходит в области глобальных ИПС сети Интернет, которая стала главной движущей силой современных технологий информационного поиска [,]. Гигантский объем доступных информационных ресурсов требует применения масштабируемых алгоритмов поиска []. Гипертексты позволяют использовать принципиально новые модели поиска, основанные на семантическом анализе коллекций документов. Высокая скорость обновления страниц, их свободное размещение и отсутствие гарантии постоянного доступа приводит к необходимости постоянного переиндексирования актуальных информационных ресурсов.

Наконец, неоднородный состав пользователей, часто не имеющих навыков работы с поисковой системой, заставляет искать эффективные способы формулировки запросов, работающие с минимальной исходной информацией [,].

4.1. Словарные информационно-поисковые системы

Словарные ИПС на сегодняшний день – самые быстрые и эффективные поисковые системы, получившие наибольшее распространение в сети Интернет. Поиск необходимой информации в словарных ИПС осуществляется по ключевым словам.

Результаты поиска формируются в ходе работы того или иного поискового алгоритма со словарем и запросом, составленным пользователем на ИПЯ.

Структура словарной ИПС (рис. Рис. 3) состоит из следующих компонентов: средства просмотра документов, интерфейса пользователя, поисковой машины, базы данных поисковых образов и индексирующего агента [,].

Рис. 3. Структура словарной информационно-поисковой системы

Информационный массив включает в себя информационные ресурсы, потенциально доступные пользователю. Сюда входят текстовые и графические документы, мультимедийная информация и т. д. Для глобальной ИПС – это вся сеть Интернет, где все документы характеризуются уникальным адресом URL[1].

Интерфейс поисковой системы определяет способ взаимодействия пользователя с ИПС. Сюда входят правила формирования запросов, механизм просмотра результатов поиска и т. д. Интерфейс поисковых систем сети Интернет обычно реали-

[1] URL – унифицированный указатель информационного ресурса (англ. Uniform Resource Locator).

зуется в среде веб-браузера. Для работы со звуковой и видео информацией применяется соответствующее программное обеспечение.

Главная функция поисковой машины – реализация принятой модели поиска. Сначала запрос пользователя, подготовленный на ИПЯ, транслируется согласно установленным правилам в формальный запрос. Затем в ходе выполнения поискового алгоритма запрос сравнивается с поисковыми образами документов из базы данных. По результатам сравнения формируется итоговый список найденных документов. Обычно он содержит название, размер, дату создания и краткую аннотацию документа, ссылку на него, а также значение меры подобия документа и запроса. Список подвергается ранжированию (упорядочению по какому-либо критерию, обычно по значению формальной релевантности).

База данных поисковых образов документов предназначена для хранения описаний индексированных документов. Структура типичной базы данных словарной ИПС подробно описана в части 1 методических указаний.

Индексирующий агент выполняет индексацию доступных документов с целью составления их поисковых образов. В локальных системах эта операция обычно осуществляется один раз: после окончания формирования массива документов вся информация индексируется и поисковые образы вносятся в базу данных. В динамическом децентрализованном информационном массиве сети Интернет применяется другой подход. Специальная программа-робот, которую называют паук (spider) или ползун (crawler), непрерывно обходит сеть. Переходы между различными документами осуществляются с помощью содержащихся в них гиперссылок. Скорость обновления сведений в базе данных поисковой системы напрямую связана со скоростью сканирования сети [, , ,]. Например, мощный индексирующий робот может обойти всю сеть Интернет за несколько недель. При каждом новом цикле обхода база данных обновляется и старые недействительные адреса удаляются.

Часть документов для поисковых машин закрыта. Это информация, доступ к которой авторизован или осуществляется не по ссылке, а по запросу из формы []. В настоящее время разрабатываются интеллектуальные методы сканирования скрытой части Интернет, но широкого распространения они пока не получили [].

Для индексирования гипертекстовых документов программы-агенты используют источники: гипертекстовые ссылки (href), заголовки (title), заглавия (H1, H2 и т. д.), аннотации, списки ключевых слов (keywords), подписи к изображениям. Для индексирования нетекстовой информации (например, файлов, передаваемых по протоколу ftp) используются URL [].

Также используются возможности полуавтоматической или ручной индексации. В первом случае администраторы оставляют сообщения о своих документах, которые индексирующий агент обрабатывает спустя некоторое время, во втором, администраторы самостоятельно вносят в базу данных ИПС необходимую информацию.

Все большее число ИПС производят полнотекстовую индексацию. В этом случае для составления поискового образа используется весь текст документа []. Форматирование, ссылки и т. д. становятся в этом случае дополнительным фактором, влияющим на значимость того или иного термина. Термин из заголовка получит больший вес, чем термин из подписи к рисунку [,].

Современные крупные ИПС должны в течение секунды обрабатывать сотни запросов. Поэтому любая задержка может привести к оттоку пользователей и, как следствие, к непопулярности системы и коммерческим неудачам. С точки зрения архитектуры, такие ИПС реализуются в виде распределенных вычислительных систем, состоящих из сотен компьютеров, расположенных по всему миру. Поисковые алгоритмы и программный код подвергаются крайне тщательной оптимизации.

В ИПС с большим объемом базы документов для ускорения их работы применяются технологии эшелонирования и прюнинга. Эшелонирование заключается в разделении базы данных на заведомо более релевантную и менее релевантную части. Сначала ИПС ищет документы по первой части базы. Если документов не найдено или найдено недостаточно, то поиск выполняется во второй части. При использовании прюнинга[1] обработка запроса автоматически прекращается после нахождения достаточного количества релевантных документов [].

Также широко применяются пороговые модели поиска, которые определяют некоторые пороговые значения для характеристик документов, выдаваемых пользователю. Например, релевантность документов обычно ограничивается некоторым значением релевантности R', например, $R' = 0{,}75$ при $0 \le R \le 1$. Вниманию пользователя предлагаются все документы со значением релевантности $R \ge R'$.

В случае ранжирования результатов поиска по дате пороговые значения определяют временной интервал даты изменения документов. Например, ИПС может автоматически отсекать документы, не изменявшиеся последние три года [].

Главным достоинством ИПС словарного типа является практически полная ее автоматизация. Система самостоятельно анализирует поисковые ресурсы, составляет и хранит их описания, производит поиск среди этих описаний. Широкий охват ресурсов сети Интернет также относится к плюсам таких систем. Значительные

[1] Pruning – англ. сокращение, удаление.

объёмы баз данных делают словарные ИПС особенно полезными для исчерпывающего поиска, сложных запросов или для локализации неясной информации.

В то же время огромное количество документов в базе данных системы часто приводит к слишком большому числу найденных документов. Это вызывает затруднения у большинства пользователей при анализе найденной информации и делает невозможным быстрый поиск. Автоматические методы индексации не могут учесть специфики конкретных документов, и количество непертинентных документов среди найденных такой системой часто бывает велико [].

Еще одним недостатком словарной ИПС является необходимость формулировать запросы к системе на специальном языке. Хотя существует тенденция к сближению ИПЯ с естественными языками, на сегодняшний день пользователь должен иметь определенные навыки в формулировании запросов.

4.2. Классификационные информационно-поисковые системы

Принцип действия классификационных ИПС заключается в распределении документов информационного массива по тематическим рубрикам. Скорость поиска в таких системах обычно невелика, однако его точность максимальна. Аналогом классификационной ИПС является любой библиотечный систематический каталог.

Иерархический классификатор поисковой системы, называемый также каталогом, определяет возможные классы, к которым могут относиться документы. Во время процесса классификации документам приписываются идентификаторы соответствующих рубрик. Эти идентификаторы и составляют поисковые образы документов, которые хранятся в базе данных системы.

Стандартные библиотечные системы классификации используются для различных целей. Во-первых, они обеспечивают удобный и предсказуемый порядок размещения документов (книг, журналов и т. п.) на полках и стеллажах, а библиографической информации – в каталогах и указателях. Кроме того, они позволяют обозначить тот или иной класс документов, а затем, в зависимости от того, получает ли пользователь в результате данного поиска слишком много или слишком мало релевантных документов, расширить или сузить этот класс, или перейти к какому-либо другому классу, связанному с этим [].

Пользователь ИПС классификационного типа сначала определяет, к какой предметной области относится интересующая его информация. Затем он выбирает соответствующую рубрику каталога. Двигаясь по иерархии рубрик, от самых общих до самых узких, в конце концов пользователь приходит к конечной рубрике, которая

содержит описания документов нужной тематической направленности. Эти документы и являются результатом поиска.

Структура классификационной ИПС показана на рис. Рис. 4. В отличие от словарной поисковой системы вместо средств индексации здесь используются средства классификации. Система каталогов является основой как процесса классификации, так и процесса поиска. Функции остальных компонентов системы аналогичны функциям соответствующих компонентов словарной ИПС.

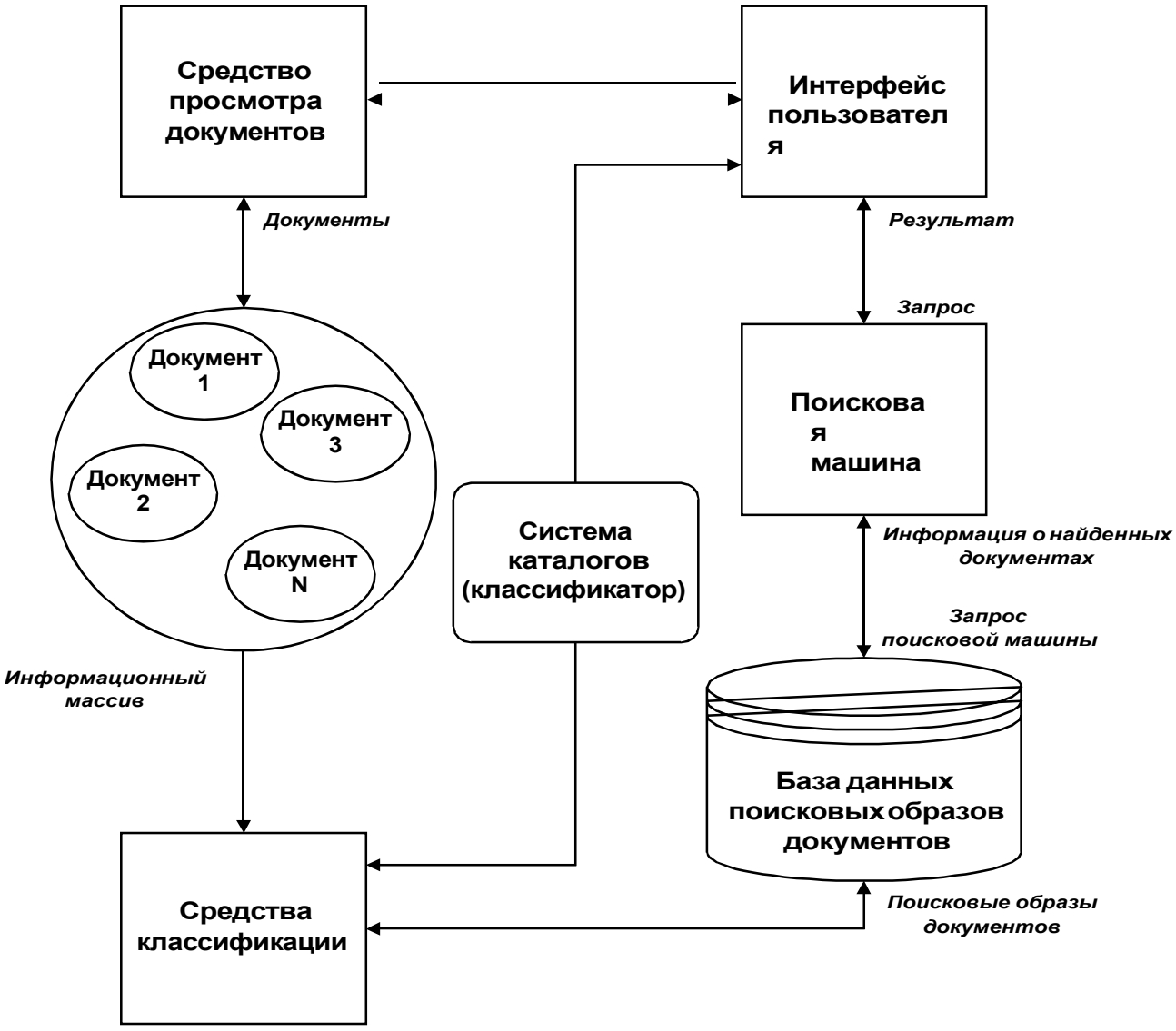


Рис. 4. Структура классификационной информационно-поисковой системы

Задача интерфейса пользователя – представить в удобном для навигации виде каталог ИПС. Обычно это реализуется с помощью иерархического списка рубрик – дерева либо через ряд связанных друг с другом гипертекстовых страниц. Интерфейс пользователя также необходим для отображения списка найденных документов.

Запрос на поиск в классификационной ИПС определяется идентификатором конечной рубрики или же последовательностью идентификаторов рубрик от верхнего

до нижнего уровня. Поисковая машина в соответствии с этим запросом обращается к базе данных и формирует список результатов поиска.

Система каталогов классификационной ИПС обычно разрабатывается людьми. Принципы и методы создания классификаторов подробно описаны в разделе 2. Эти положения применимы и при автоматической генерации классификаторов для некоторого множества документов. Однако полученные таким образом системы классификации трудны для восприятия массового пользователя и не всегда обеспечивают адекватное распределение документов.

Аналогичные трудности встречаются и при использовании средств классификации, которые также могут быть ручными и автоматическими. В условиях непрерывного роста объема информации автоматическое распределение документов по сравнительно небольшому фиксированному набору каталогов приводит к тому, что число документов в конечных рубриках резко увеличивается. Эффективность поиска снижается, так как пользователь, находясь в конечном разделе классификатора, не может повысить точность, сужая число выдаваемых документов. Если динамически добавлять в классификатор новые разделы, то рано или поздно его структура станет настолько сложной, что использовать систему будет невозможно [].

В то же время проведенная коллективом специалистов систематизация документов обеспечивает предсказуемый с точки зрения пользователей порядок размещения документов в каталоге [].

Ручная разработка классификатора определяет один из главных недостатков классификационных ИПС. Различные области человеческой деятельности и знаний могут получить разную оценку своей относительной важности у разных групп разработчиков. Глубина проработки и ширина охвата того или иного раздела классификатора часто бывает разной в различных ИПС, что обуславливает трудности перехода пользователей от одной системы к другой.

Описанная проблема свойственна как глобальным, так и локальным ИПС классификационного типа. Если пользователь не имеет четкого представления об интересующей его предметной области, найти необходимые документы будет довольно трудно. Однако при этом классификатор может помочь сформулировать информационную потребность или расширить запрос синонимическими терминами [].

Ручные методы составления классификаторов и распределения по ним документов занимают по сравнению с автоматическими гораздо больше времени и имеют гораздо большую стоимость. Это оправдывает себя лишь в небольших локальных поисковых системах. Глобальные ИПС сети Интернет в состоянии классифицировать только крайне малую часть всех документов сети. Поэтому основное досто-

инство классификационных ИПС заключается в качестве предоставляемой ими информации. Просмотренные людьми и систематизированные документы позволяют достигать высокой точности поиска [, ,].

4.3. Метапоисковые системы

Любая поисковая система имеет собственный информационный массив, который состоит из множества доступных для поиска документов. Это множество документов всегда ограниченно. Локальные поисковые системы по определению работают с некоторым фиксированным объемом информационных объектов. Число документов в сети Интернет постоянно растет, однако скорость увеличения числа доступных для поиска документов всегда меньше скорости их появления в сети.

В настоящее время ни одна ИПС не может охватить все ресурсы в Интернет. Поэтому поиск с использованием какой-либо одной ИПС часто не может полностью удовлетворить информационную потребность пользователя. В такой ситуации приходится повторять один и тот же запрос в нескольких поисковых системах. Для увеличения широты охвата и расширения возможностей поиска, а также для облегчения работы пользователей были разработаны так называемые метапоисковые системы.

Метапоисковые системы не имеют собственных баз данных поисковых образов документов, средств индексации и классификации. При поиске они используют ресурсы других поисковых систем [,]. За счет одновременного обращения к взаимно дополняющим друг друга базам данных нескольких ИПС в метапоисковых системах достигаются максимальные значения полноты поиска [].

Порядок работы с метапоисковой системой, структура которой представлена на рис. Рис. 5, можно упрощенно описать следующим образом. Пользователь в соответствии со своей информационной потребностью составляет запрос на поиск. Метапоисковая система передает этот запрос другим ИПС, которые и осуществляют поиск по своим информационным массивам. Затем результаты поиска в виде списков найденных документов от различных ИПС поступают обратно в метапоисковую систему, и в ней формируется итоговый список документов, который предлагается вниманию пользователя. Найденные документы ранжируются в порядке их следования в результатах поиска каждой из ИПС. При этом существенно повышается релевантность тех документов, которые были одновременно найдены в нескольких ИПС.

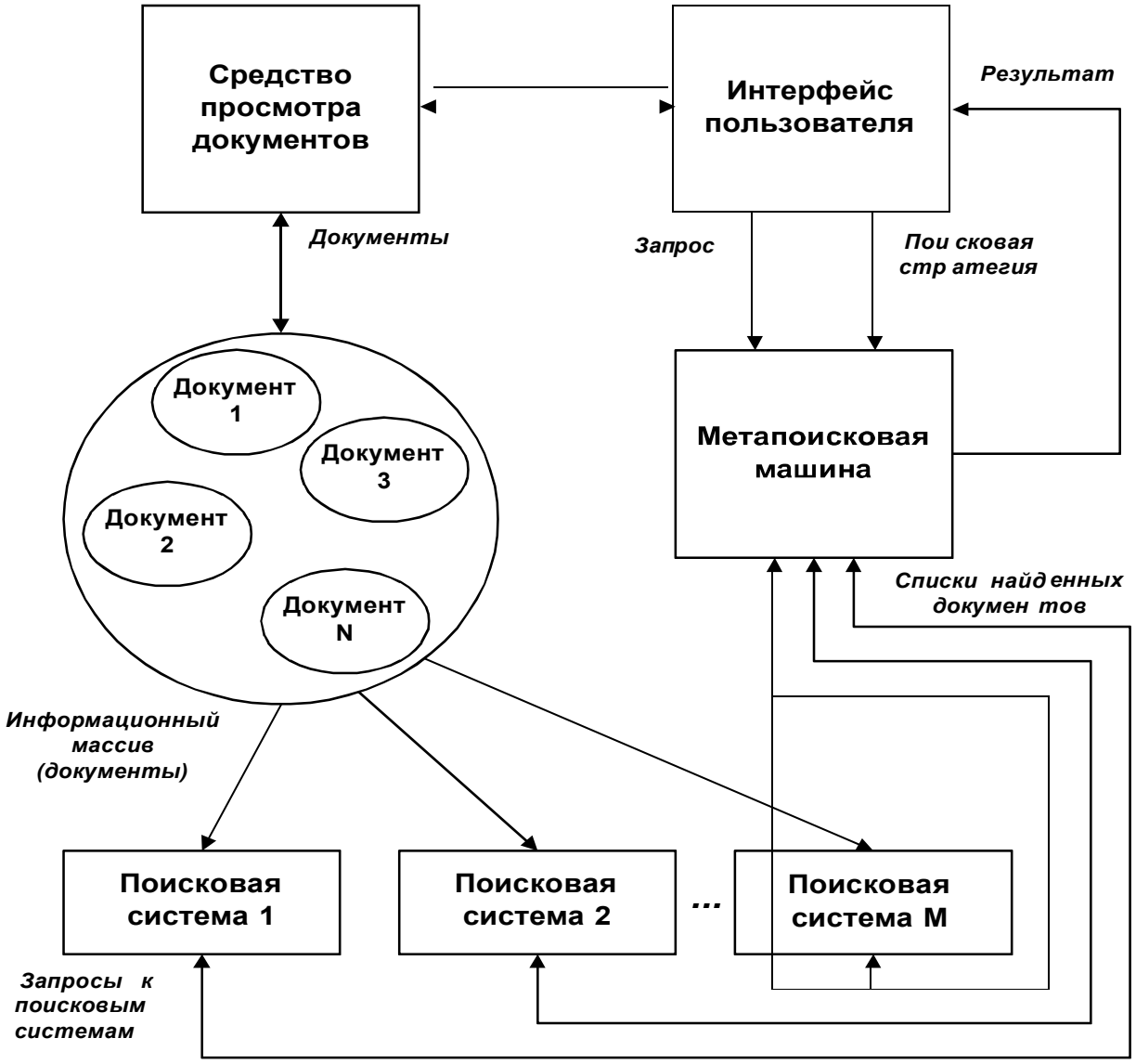


Рис. 5. Структура метапоисковой системы

Главная проблема, связанная с реализацией данного алгоритма, заключается в том, что поисковые системы используют разные методы индексации, имеют различные информационные массивы и, как следствие, базы индексированных документов различной полноты. Поэтому запрошенная пользователем информация может быть найдена в одной системе и не найдена в другой. В этом случае можно получить несколько полностью релевантных документов от одной ИПС, которые будут перемешаны с частично релевантными документами из другой (например, в случае частичного совпадения документа и запроса) [,].

Современные метапоисковые системы позволяют преодолеть эти трудности. Во-первых, каждая ИПС придерживается (в течение достаточно долгого времени) собственных правил ранжирования результатов поиска, что используется метапоисковой машиной при определении релевантности документов, полученных от разных систем. На значение релевантности также влияет рейтинг ИПС, определяемый каче-

ством поиска в ней, и общее количество документов, найденных по запросу (это также позволяет оценить полноту базы поисковых образов конкретной ИПС) [].

Наконец, главный метод корректного ранжирования заключается в статистическом анализе результатов поиска в различных системах. Обычно результаты поиска содержат названия (заголовки) и краткие описания (аннотации) найденных документов. Метапоисковая машина определяет частоты встречаемости терминов запроса в заголовках и аннотациях документов и присваивает каждому документу некоторый вес, используемый затем при ранжировании. Подобная обработка позволяет не только понижать ранг документов, в описании которых вообще нет ключевых слов, как потенциально нерелевантных запросу, но и находить строгое соответствие в том случае, если все ключевые слова встречаются в описании документа.

На схеме (рис. Рис. 5) пользователь помимо запроса к поисковой системе определяет стратегию поиска. Формирование стратегии поиска предполагает выбор пользователем типа информационных объектов, которые нужно найти с помощью ИПС (файлы, новостные сообщения, гипертекстовые документы и др.), выбор области поиска (русскоязычная часть Интернета, англоязычная часть или глобальный поиск по всей всемирной сети), а также выбор ИПС, к которым должна обращаться во время поиска метапоисковая система. В результате объединения текста запроса на ИПЯ и ряда поисковых предписаний формируется так называемый расширенный запрос, который затем ретранслируется метапоисковой машиной другим ИПС.

Заключение

Практика информационного поиска ставит перед исследователями все новые и новые задачи, не позволяя останавливаться на достигнутом и заставляя создавать новые теории и методы, проектировать и моделировать новые системы.

Распознавание, индексирование и поиск документов разных форматов и представлений, включая мультимедийные; использование инструментов и ресурсов систем управления базами данных в ИПС; развитие технологий поиска в сети Интернет, в частности, обработка очень больших объемов информации, архитектурная реализация систем – вот лишь самые основные перспективные направления развития современных ИПС [,].

Следует также сказать о постоянном повышении требований к ИПС в плане эффективности организации взаимодействия человека и поисковой системы, главным образом, к проектированию пользовательских интерфейсов ИПС, ориентации систем на манеру работы человека, его ожидания и предпочтения.

Библиографический список

1.Добрынин В. Ю. Теория информационно-логических систем. Информационный поиск: Метод. указания к курсу информационного поиска. – СПб. : Изд-во СПбГУ, 2012.

2.Дубинский А. Г. Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. – 2011. – № 4.

3.Капустин В. А. Основы поиска информации в Интернете. Методическое пособие. – СПб. : Институт "Открытое общество", С.-Петерб. отд-ние, 2011.

4.Когаловский М. Р. Перспективные технологии информационных систем. М. : ДМК Пресс : Компания АйТи, 2013.

5.Кромер В. В. Об одной поправке к каноническому закону // Телеконференция "Информационные технологии в гуманитарных науках". – Казань, 2010.

6.Кураленок И. Е. Оценка систем текстового поиска / И. Е. Кураленок, И. С. Некрестьянов // Программирование. – 2012. – № 4.

7.Некрестьянов И. С. Системы текстового поиска для Веб / И. С. Некрестьянов, Н. Пантелеева // Программирование. – 2012. – № 4.

8.Попов А. Поиск в Интернете – внутри и снаружи // Internet. – 2013. – № 2.

9.Сегалович И. В. Как работают поисковые системы // Мир Internet. – 2012. – № 10.