Лекция 4. Дескрипторный поиск

(Нужно перенумеровать)

На прошлой лекции мы рассмотрели два типа информационно-поисковых языков: язык библиографических данных и классификационные языки. Первые из них описывают документ с «внешней» стороны. Они применяются для поиска документов в алфавитных **библиографических** каталогах, организованных по выходным данным документов, образованным в процессе ***создания*** документа. Языки второго типа описывают содержание документа путём отнесения его к какой-либо отрасли знания. Они применяются для поиска документов в **систематических** каталогах, организованных по индексам классификационных систем, присвоенных документу в процессе упорядочения уже созданных документов в документальных фондах ***в хранилищах***. В каждой конкретной информационной системе (библиотеке, информационном органе) эти языки могут быть устроены по-разному. В разных случаях могут быть выбраны разные поисковые элементы библиографических данных, могут быть выбраны разные системы тематической классификации. Мы подробно рассмотрели Универсальную десятичную классификацию (УДК), познакомились с Государственным рубрикатором научно-технической информации (ГРНТИ), упомянули и другие библиографические классификации (ББК, ДКД). Все они образуют различные информационно-поисковые языки классификационного типа.

1. Языки предметных рубрик

Кроме библиографических (алфавитных) и систематических каталогов в библиотеках распространены **предметные** каталоги. Сущность предметного каталога заключается в том, что содержание документа кратко формулируется при помощи одного или нескольких типовых ключевых слов, получивших название **предметных рубрик**. Затем предметные рубрики располагаются по алфавиту и под каждой из них как под заголовком[1] собираются библиографические описания документов. Составление предметных рубрик и распределение по ним документов называется **предметизацией**. Задача предметизации обычно состоит в том, чтобы указать главный объект рассмотрения в документе и, может быть, основные его аспекты и основные отношения к другим предметам.

В отличие от библиографических классификаций предметизация распределяет документы по предметам или понятиям, не соотнося их с какими-либо областями знания. Это различие делает классификационный и предметизационный принципы организации документов независимыми,

[1] Предметные рубрики также называются **предметными заголовками** (subject headings), и этот вариант термина лучше отражает природу термина, выбранного для обозначения содержания ряда документов в каталоге. А собственно **предметными рубриками** следует называть совокупности документов, объединенные одним предметным заголовком.

дополняющими друг друга, предназначенными для поиска документов по разным типам запросов. Предметизация даёт возможность собирать в одном месте документы по таким комплексам как то: конкретный материал, свойство, изделие, явление природы или общества, род деятельности, географическое понятие и т. п., собирая под каждым предметным заголовком весь массив знаний, безотносительно к тому, какой области науки эти знания принадлежат.

Другим отличием предметизации от классификации является то, что заранее составленный список предметных заголовков не ограничивает подробности анализа содержания документа. Если документ посвящён вопросу, не отражённому в списке предметных заголовков, всегда имеется возможность сформулировать новый заголовок самостоятельно. Обычно же при выборе предметной рубрики для документа руководствуются заранее составленным списком предметных заголовков. Но не представляет труда внести в него вновь образованную предметную рубрику, которая займёт своё надлежащее место, определяемое алфавитным порядком. Это обычно не удаётся делать в языках классификационного типа, где введение новых классов зачастую влечёт преобразование большой части классификационных связей.

У каждого предметного заголовка могут в принципе быть подзаголовки, делящие документы в рубрике на подрубрики. В некоторых случаях к предметному заголовку могут быть даны ссылки на другие рубрики, где могут находиться документы по сходному предмету. Таким образом, в списке предметных заголовков одна запись может иметь довольно сложный характер. Вот пример фрагмента словаря предметных рубрик:

…
самовары
самолёты
- военные
- гражданские
- - грузовые
- - магистральные
- - пассажирские
см. также авиалайнеры
саморезы
см. винты-саморезы
самоходные баржи
см. теплоходы грузовые речные
…

Традиционная каталожная техника не позволяет раскрывать содержание документа предметными рубриками с достаточной полнотой. Предметные заголовки отражают только основной предмет документа, и даже документы многопланового, обзорного характера могут быть отражены в каталоге лишь в ограниченном числе предметных рубрик. Недостаток места, трудоёмкость составления картотечных каталогов заставляли библиографов разрабатывать довольно сложные правила оптимального

выбора используемых предметных заголовков[2]. Эти правила вместе со словарями рекомендуемых предметных заголовков составляют информационно-поисковые **языки предметных рубрик (ИПЯ предметных рубрик, предметный ИПЯ).**

Языки предметных рубрик используются не только как основа для предметного каталога документов, но также и как вспомогательное средство для пользователя систематическим каталогом, основанным на тематической классификации. В тех случаях, когда пользователь знает предмет своего интереса, но не знает, к какой отрасли знания он относится, он может обратиться к **алфавитно-предметному указателю**, в котором для каждой предметной рубрики указывают подходящий раздел (разделы) тематической классификации, в котором собраны знания по данному предмету. В этом случае мы имеем дело с языком предметных рубрик в функции поиска не документов, а тематических разделов систематического каталога, или, что то же самое, для поиска классов тематической классификации.

2. Языки ключевых слов

Современная компьютерная техника снимает ограничения по объёму каталогов и снижает трудоёмкость их составления. Поэтому получила распространение идея приписывать документам **все** ключевые слова, используемые в документе, и в электронном каталоге иметь инверсный файл записи адресов документов, использовавших **каждое** ключевое слово.

Под **ключевыми словами** в данном случае понимаются наиболее существенные для выражения содержания документа полнозначные слова и словосочетания, обладающие назывной (номинативной) функцией. Поиск документа при этом должен происходить, как правило, не по одному ключевому слову (не по одной предметной рубрике, как в случае языка предметных рубрик), а по формулировке поисковой потребности, содержащей ряд ключевых слов, полно описывающих тему поиска. В процессе поиска по записям инверсного файла, соответствующим ключевым словам запроса, должны производиться логические операции над множествами адресов документов аналогично тому, как это делается для сложных запросов на языке библиографических данных.

Ключевые слова образуют новый способ описания и поиска документов – **язык ключевых слов (ИПЯ ключевых слов),** который сливается с языком библиографических данных в единый программный комплекс с едиными процедурами поиска.

Описанная организация языка ключевых слов позволяет сильно экономить на ресурсах памяти по сравнению с языком предметных рубрик. Так, если, допустим, документ заиндексирован шестью терминами, то мы имеем возможность находить его по всем сочетаниям этих терминов. А в

---

[2] Подробнее см. А.И.Михайлов, А.И.Чёрный, Р.С.Гиляревский. Основы информатики. – М.: Наука, 1968. – 756 с. // с. 346 – 366 Алфавитно-предметные классификации.

языке предметных рубрик мы должны были бы иметь порядка 100 предметных заголовков вместо 6 ключевых слов, чтобы точно указать тематику этого документа. Примером может служить предыдущая лекция. Её содержание можно описать следующими шестью ключевыми словами:

ИПЯ, Библиографические данные, Библиографические классификации, УДК, ББК, ГРНТИ

В языке предметных рубрик только такие предметные рубрики будут давать точное описание этой лекции, которые будут содержать все эти ключевые слова в качестве заголовка или подзаголовков. Но сочетаться эти слова могут в разном числе и в произвольном порядке, что составляет более 1000 вариантов. Среди этих вариантов меньшая часть является осмысленной формулировкой темы, и потому реально необходимо включать в словарь меньше рубрик, но всё равно их число намного превосходит число ключевых слов.

3. Информационно-поисковый тезаурус

При всём своём удобстве язык ключевых слов обладает очевидными недостатками. Так, на запрос «поваренная соль» не будут выданы документы, заиндексированные терминами «хлористый натрий» и «натрия хлорид» (все три термина обозначают одно и то же вещество). На запрос «языкознание» не будут выданы документы о «лингвистике» и «языковедении» (одно и то же), а также документы о разделах этой науки: «грамматика», «фонетика», «орфография» и др. Зато на запрос «стабилизаторы» будут выданы документы как о стабилизаторах в хвостовом оперении самолётов, так и о стабилизаторах электрического тока и стабилизаторах смесей химических веществ, хотя пользователю были нужны документы только по одной из этих областей знания. Более того, на запрос «калия хлорид» будет выдан документ, в котором идёт речь о бромидах калия и хлоридах натрия, но не о хлоридах калия, поскольку в поисковом образе будут представлены отдельные ключевые слова – «калий», «натрий», «хлорид», «бромид» без указания связи между ними.

Для преодоления этих недостатков нужно, чтобы поисковая система реагировала не на формальное совпадение слов, а на совпадение **понятий**, а ещё лучше – на совпадение **смысла** текстов.

Понятие может быть задано как совокупность всех способов обозначения нго в текстах, т. е. как множество **синонимов**. Для этого в автоматизированную систему должен быть введён словарь, в котором было бы указано, какие слова и выражения обозначают одно и то же понятие:

Библиографическое описание = бибописание
Компакт-диск = CD
Компьютер = вычислительная машина = ЭВМ
Лазер = оптический квантовый генератор
Натрия хлорид = хлористый натрий = поваренная соль
Пневмония = воспаление лёгких
Ядерная энергия = атомная энергия
Языкознание = лингвистика = языковедение

При этом одно из эквивалентных выражений выбирается как нормативное слово, допущенное для использования, а остальные синонимы служат справочным материалом и подлежат замене на нормативное обозначение понятия.

Идею – использовать при индексировании документов и запросов такие словари синонимов – предложил американский математик К. Муэрз (C. Mooers) ещё в 1947 г. для системы механического поиска библиографических карточек. Совокупность синонимов, обозначающих одно понятие он назвал **«дескриптор»** («описатель»). Расширительно дескриптором называют также тот нормативный синоним, который в системе заменяет собою остальные синонимы. Последние получили наименование **«аскрипторы»** («не описатели, не писомые»).

В дальнейшем было предложено вводить в дескрипторный словарь другие отношения – отношения дескрипторов как целое, а сам словарь получил название **информационно-поискового тезауруса**. Словом «тезаурус» ранее именовали списки слов, охватывающие всё лексическое **богатство** какого-либо источника (Библии, Гомера, например). Само по себе слово «тезаурус» в греческом языке (θεσαυρός) означает «сокровище». А в словаре ключевых слов (дескрипторов) содержались **все** слова, допущенные для использования при индексировании, т.е. всё лексическое сокровище системы.

Информационно-поисковый тезаурус (ИПТ) – это словарь терминов определённой области знания, в котором между терминами зафиксированы путём ссылок смысловые связи понятий, отражающие взаимодействие (отношения) объектов и явлений действительности. Пользуясь философским языком, можно сказать, что тезаурус отражает **онтологию** предметной области. Поэтому в последнее время распространилось наименование этого понятия термином «онтология»[3]. Системы описания документов и запросов с помощью дескрипторов и информационно-поисковых тезаурусов называют информационно-поисковыми **языками дескрипторного типа,** или **дескрипторными ИПЯ**.

4. Координатное индексирование

Функционирование информационной системы с информационно-поисковым языком дескрипторного типа происходит следующим образом. Документ, поступающий в хранилище информационной системы, подвергается анализу на содержание в нём терминов, включённых в тезаурус. Термин обнаруженный в документе, приписывается ему аналогично классификационному индексу. Но если при индексировании документов с помощью языка классификационного типа мы в норме ограничиваемся

[3] Лица, введшие этот термин в употребление, пришли к необходимости использовать такой словарь самостоятельно, не имея знаний о предыдущем опыте разработки информационных систем, в ходе которого впервые и была выдвинута концепция ИПТ. Поэтому, когда им потребовался термин для обозначения соответствующего понятия, они не могли воспользоваться уже имеющимся выражением.

отнесением документа к одной классификационной рубрике каталога, то при сравнении с ИПТ мы в идеале должны стремиться проверить документ на вхождение каждого термина из тезауруса. Классификационным языком мы характеризуем документ его местом в линейном ряду классификационных рубрик, а дескрипторы тезауруса характеризуют его с разных сторон, аналогично тому, как географические координаты характеризуют местоположение не только по отношению север-юг, но также по направлению восток-запад, и ещё по направлению верх-низ. Из этой аналогии происходит наименование принципа индексирования документов дескрипторами тезауруса – **координатное индексирование**. Впрочем координатное индексирование может быть реализовано и с помощью классификационной системы, как мы это видели на примере использования УДК (один документ может быть отнесён к двум и более классам). Но идеал классификационного индексирования – определить единственное «правильное» место документа в каталоге, а идеал дескрипторного индексирования – определить отношение документа ко всем включённым в ИПТ терминам (координатам).

Термин может характеризовать документ не только простой констатацией своего присутствия-отсутствия, но также и важностью, **весом** термина в документе (аналогично численным значениям географических координат). Веса терминов могут вычисляться по числу их употребления в данном документе, а также назначаться по вхождению в заглавие и другие важные элементы документа. Разные слова могут также иметь собственные внутренние особенности, влияющие на их «дескрипторроспособность». Сильно облегчают анализ документов списки **«запрещённых»** слов, куда входят слова грамматического характера, а также слова со слишком неопределённым значением, которые заведомо не могут служить для описания тематики документа и которые не входят в число дескрипторов тезауруса. (Такой список ещё называют **стоп-словарь**). Веса дескрипторам также могут присваиваться путём экспертного анализа их смысловой важности для данного документа.

При анализе документа можно также указать на **связь** в нём отдельных дескрипторов. Ясно, что от связи слов весьма сильно зависит содержание документа. Так, одно дело, когда в документе идёт речь, например, о производстве ***пива*** в странах ***Балтики***, а другое – когда о производстве ***пива «Балтика»***. Различие этих документов можно отразить, указав, что во втором случае слова ***пиво*** и ***Балтика*** употребляются (почти) всегда рядом. Можно указывать, что слова употребляются в документе в одном предложении, в одном абзаце. Можно указывать расстояние между словами в числе промежуточных слов, что в некоторой степени отражает их смысловую связь. Применяются и другие **указатели связи.**

При интеллектуальном анализе документа можно дескрипторам приписывать их **смысловую роль** в документе, например:

«*главный предмет рассмотрения*»,

«*характеристика предмета*»,
«*цель исследования*»,
«*результат*»,
«*условия*»,
и тому подобное.

Такие пометы при дескрипторах имеют название **указатели роли.**

5. Поисковые образы

В результате индексирования документа на входе в систему он приобретает своё описание в виде перечня ключевых слов (дескрипторов), которые могут дополняться их весами, связями и указателями роли. По этому описанию внутри системы составляются каталоги, служащие для поиска документов и выдачи их из хранилищ. Соответственно эти описания называют «**поисковый образ документа**», или в специальной литературе сокращением «**ПОД»**. Составление поискового образа, вообще говоря, представляет собой непростую интеллектуальную задачу. Её сложность можно понять, ознакомившись со стандартами на данную работу: ГОСТ 7.66 (координатное индексирование документов) и ГОСТ 7.52 (поисковый образ на машинных носителях). В полном объёме эту работу можно осуществить когда речь идёт об ограниченном потоке данных, допустим о поступлениях новых специальных документов в архив или в библиотеку предприятия с узкой сферой деятельности. Такие системы интеллектуального индексирования получили некоторое распространение ещё в середине прошлого века (60 – 70 годы).

Как нетрудно видеть, процедура выявления дескрипторов в документе может производиться автоматически, путем формального сравнения текста документа со словарём. Хотя тут тоже есть свои проблемы. Необходимо разработать довольно сложные программы идентификации одного и того же слова в разных грамматических формах. Нужно опознавать словосочетания как формы выражения одного понятия, нужно как-то решать проблемы многозначности слов и синонимии. А кроме того автоматическое индексирование может быть применено только к документам на машинных носителях. Но с развитием электронного документооборота, электронной полиграфии, а особенно с распространением Интернета электронные документы становятся преобладающими носителями информации, и автоматизация индексирования становится необходимым условием функционирования информационных систем.

Так или иначе, внутри информационной системы документ предстаёт своим поисковым образом. Запрос пользователя на поиск информации также должен быть представлен в той же форме, что и документ. Тогда сравнение поисковых образов документов с **поисковым образом запроса (ПОЗ)** даст ответ на соответствие документа запросу. ПОД и ПОЗ одинаково представляются как перечни дескрипторов с возможным указанием весов, ролей и связей дескрипторов. Однако для их сравнения нужно задать ещё

критерий смыслового соответствия. До какой степени ПОЗ должен совпадать с ПОДом? Если в запросе задано только одно понятие, а в ПОДе кроме него имеется ещё и другое, то следует ли такой документ выдавать? А если наоборот? В разных случаях информационной потребности пользователя могут отвечать разные критерии смыслового соответствия. Совокупность поискового образа запроса и критерия смыслового соответствия называется **поисковым предписанием**. Оно представляет информационную потребность пользователя внутри системы и управляет выдачей документов из неё.

6. Дескриторные ИПЯ

В целом дескрипторный информационно-поисковый язык представляет собой сложный комплекс, который включает:

- информационно-поисковый тезаурус, представляющий внутри информационной системы модель онтологии предметной области поиска;
- правила представления в поисковых образах предмета информационного поиска как фрагмента онтологической модели, заданной в тезаурусе;
- массив поисковых образов документов, где каждый ПОД представляет собой модель понятийного содержания документа, соотнесённую с предметной областью через её модель, заданную в тезаурусе;
- правила формирования поискового предписания, выражающего информационную потребность пользователя дескрипторами тезауруса;
- правила определения критерия смыслового соответствия ПОД и ПОЗ, отвечающего информационной потребности пользователя.

Этот набор компонентов дескрипторных ИПЯ позволяет рассматривать их как потенциально мощный инструмент смыслового поиска информации. Реальная эффективность этого инструмента определяется степенью развития, степенью совершенства каждого из пяти компонентов. К сожалению в настоящее время имеется опыт разработки и эксплуатации только довольно примитивных дескртипторных языков. Однако в нормативных документах, разработанных в конце прошлого века, уже были намечены пути их развития, которое предстоит осуществить теперь на базе новых компьютерных технологий. Важнейшими из этих нормативных документов являются следующие международные и государственные стандарты, с которыми студентам необходимо ознакомиться.

Государственные стандарты России:

- ГОСТ 7.24-90 Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению

- ГОСТ 7.25-2001 Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления
- ГОСТ 7.47-84 Коммуникативный формат для словарей информационных языков и терминологических данных. Содержание записи
- ГОСТ 7.52-85 Коммуникативный формат для обмена библиографическими данными. Поисковый образ документа
- ГОСТ 7.66-92 Индексирование документов. Общие требования к координатному индексированию
- ГОСТ 7.74-96 Информационно-поисковые языки. Термины и определения

Международные стандарты:

- ИСО 2788:1986 Документация. Руководство по построению и разработке одноязычных тезаурусов
- ИСО 5963:1985 Документация. Методы анализа документов, определения их темы и подбора индексирующих терминов
- ИСО 5964:1985 Документация. Руководство по построению и разработке многоязычных тезаурусов
- ИСО 6156:1987 Формат для обмена терминологическими и/или лексикографическими данными (MATER)