5 лекция. Информационно-поисковый тезаурус

На прошлой лекции мы остановились на понятии информационно-поискового языка дескрипторного типа. Его основой служит информационно-поисковый тезаурус. Рассмотрим эти понятия подробнее.

Информационно-поисковый тезаурус (ИПТ) – это словарь терминов определённой области знания, между которыми (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». Лица, пользующиеся этим словом (введшие его в употребление) пришли к необходимости использовать такой инструмент в информационной работе самостоятельно, не имея знаний о предыдущем опыте разработки информационных систем, в ходе которого впервые и была выдвинута концепция ИПТ. Автор этой концепции – американский разработчик библиотечных информационных систем К. Муэрс (C. Mooers). Он предложил в словарях ключевых слов указывать отношения синонимии, рассматривая КС, соединённые этим отношением как единый элемент смыслового индексирования документов – дескриптор.

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

1. 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 языком классификационного типа в норме мы ограничиваемся отнесением документа к одной классификационной рубрике каталога, то при сравнении с ИПТ мы должны стремиться в идеале 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 «дескрипторноспособность». Так для облегчения входного анализа документов составляют 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 (Эти списки ещё называют **стоп-словарь**). Веса дескрипторам могут присваиваться путём интеллектуального анализа их смысловой важности для 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 ***балтийских*** странах, а другое – когда о производстве ***пива «Балтийское»***. Различие этих документов можно отразить, указав, что во втором случае слова ***пиво*** и ***балтийское*** 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. При интеллектуальном анализе документа можно дескрипторам приписывать их смысловую роль в документе: то ли данный дескриптор означает «*главный предмет рассмотрения*», то ли его «*характеристику*», то ли «*цель исследования*», то ли «*результат*», то ли «*условия*», и тому подобное. Такие пометы при дескрипторах имеют название **указатели роли.**

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 ГОСТ7.52 (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, а автоматизация индексирования – 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.

При автоматическом выявлении дескрипторов в документе и при возможностях современных компьютеров можно отказаться от ограничения определённым кругом слов в тезаурусе, а включать в поисковый образ документа **все слова подряд** (может быть за исключением слов из стоп-словаря). Именно так поступают современные поисковые машины Интернета. Тогда спрашивается, зачем нужен нам тезаурус как словарь

терминов, допускаемых для составления поисковых образов документа и запроса? Вот разработчики поисковых машин и решили, что он им не нужен. Более того, не имея словаря, поисковые машины не могут опознавать термины, представленные словосочетаниями, распознавать многозначные и синонимичные слова. В результате получаем то, что можно назвать **языком пословного индексирования.**

Простейшие поисковые программы при этом не занимаются даже отождествлением различных грамматических форм одного слова. «Робот» этих «поисковых машин» («наук») постоянно просматривает всю сеть WWW, расчленяет каждый сайт на отдельные словоформы «от пробела до пробела», удаляет неинформативные «стоп-слова», а на остальные записывает в свой внутренний словарь с указанием для каждого слова адресов страниц, на которых это слово употреблено. Да! Этих слов – около 100 тысяч (для одного языка). Да! У каждого слова – миллион адресов. Современная компьютерная техника позволяет запоминать такие объёмы данных и эффективно их просматривать. Общий объём такого словаря можно оценить таким образом: 100 тысяч слов х 1 млн. адресов х 100 символов в каждом адресе. Это составляет 10^{13} символов, т.е. 10 млн. мегабайт = 10 тысяч терабайт. Этот объём примерно сравним с объёмом данных, обрабатываемых, например, американской системой наблюдения за Землей в течение 10 дней. Так вот, пауки поисковых машин тоже примерно за 10 дней обегают всю сеть WWW, и сведения о ваших сайтах становятся известны системе примерно через это время.

Далее, имея такой «индексный файл», поисковая машина на ваш запрос мгновенно выдаёт списки адресов, зафиксированных при каждой словоформе запроса, производя над ними операцию пересечения множеств или другие, если это указано пользователем через заполнение специального формуляра сложных запросов.

Развитые поисковые машины (а с течением времени они всё больше становятся развитыми), они при этом учитывают грамматику естественного языка (Яндекс – русскую грамматику) и в своём индексном файле объединяют записи словоформ, относящиеся к одной лексеме (одному слову с одним значением, но в разных формах, склонения или спряжения). Поэтому будет одинаковой выдача, например, на такие два запроса:

«информационно-поисковые языки»

и «информационный поисковый язык»

Это позволяет находить такие тексты, где предмет запроса не только ***называется***, но также и те, где он присутствует в косвенных падежах как объект или обстоятельство каких-либо действий. С другой стороны это позволяет сокращать индексный файл за счёт объединения записей словоформ.

Яндекс также вычисляет ***веса*** слов в документе (на WWW – странице) и при выдаче ранжирует документы по «релевантности» - по степени

соответствия документа запросу. Сначала выдаются ссылки на страницы, где суммарный вес запрошенных слов наибольший. Конкретная методика грамматического анализа и вычисления весов – это фирменное «know haw» и не разглашается. Сохраняются также сведения о совместной встрече слов в предложениях и абзацах

3. Но всё же, например, при запросе всех документов по **лингвистике** вам не будут выданы документы, где говорится о **языкознании** или **языковедении** (это синонимы). Вы не найдёте документов о **грамматике, лексикологии, фонетике** и других понятиях, входящих в объём запрошенного понятия. Вам конечно не будут выданы документы о **семиотике** или **информатике**, которые, хотя и не входят в объём понятия **лингвистика**, но тесно с ним связаны и могут представлять для вас интерес, если уж вас интересует «всё о лингвистике».

Одним словом: Нам обычно требуется найти документы не те, в которых употреблено некоторое ***слово***, а те, в которых рассматривается соответствующее ***понятие*** или соответствующий **объект**. Т.е. поиск должен идти не по словам (***лексический поиск***), а по их смыслу (***семантический поиск)***. Специалисты в области научной и технической информации в принципе решили эту задачу ещё в середине прошлого века, предложив концепцию информационно-поискового тезауруса. Они пытались её реализовать сперва даже на ручных каталожных карточках, затем на механических сортировальных машинах, и наконец – на больших вычислительных машинах конца XX века (main frame computers). Здесь были достигнуты обнадёживающие успехи. В нашей стране действовало до 200 поисковых систем в различных областях знания, использовавших тезаурусы. Но тут произошла компьютерная революция – появились персональные компьютеры и Интернет, и сменилось поколение машин, разработчиков и общий подход к проблеме. В нашей стране это ещё усугубилось преобразованиями в общественной жизни, и в результате от прежних достижений почти ничего не сохранилось, кроме идей. А в идейном плане наша страна (как это обычно) шла впереди западной науки. Но Запад шёл впереди по технике. А технический прогресс последнее время шёл настолько быстро, что оказалось просто дешевле добиваться приемлемых характеристик информационных систем за счёт механического наращивания их быстродействия т объёмов памяти. И только в самое последнее время идея поиска информации ***по смыслу*** вновь возродилась под наименованием «семантический вэб» и «онтологии».

Ну что ж! Онтологии – так онтологии. Можно называть информационно-поисковые тезаурусы и онтологиями. Хоть это и не правильно. Зачем употреблять слово не в своём прямом значении – «философское учение о бытии, обо всём сущем»?

4. Обратимся теперь к тому, как эта онтология представляется в виде информационно-поискового тезауруса (ИПТ). Правила разработки и форма представления ИПТ была определена стандартом ГОСТ 7.25 и стандартами

других стран (США) в начале 70х годов. Форма представления ИПТ на машиночитаемых носителях, оптимальная для доинтернетовской эпохи была зафиксирована в 80х годах отечественным стандартом ГОСТ 7.47 и сходными зарубежными международными стандартами. Форма представления «онтологий», т.е. информационно-поисковых тезаурусов нашего времени определяется разработкой в рамках профессиональной ассоциации С3W (которая развивает стандарты Интернета), разработкой языка OWL (Ontology Web Language), к5оторый становится международным стандартом де факто.

Итак, ИПТ – это словарь терминов, в котором прежде всего указаны ссылки от терминов к их синонимам (эквивалентам), к обобщающим (родовым) и к частным (видовым) понятиям. При наличии в системе такого словаря она имеет возможность в ответ на запрос с термином, допустим, **языкознание** присоединить в выдачу документы, соответствующие терминам синонимам: ***лингвистика*** и ***языковедение***. Обычно присоединяются также документы по видовым понятиям: ***грамматика, лексикология, фонетика*** и др. При специальном указании пользователя об исчерпывающем сборе информации сюда могут быть добавлены также документы по терминам обобщающих понятий (***филология*** – включает также литературоведение) и ассоциативных понятий (***семиотика, информатика*** и др.).

Если в тезаурус включить другие смысловые отношения терминов, например – часть – целое, причина – следствие, свойство – носитель, процесс – инструмент и т.п., то открывается возможность точно формулировать поисковые образы довольно сложных логически запросов типа:

«Найти документы, в которых по причине А объект Б является носителем свойства В»

По таким запросам будут получены вполне вразумительные данные

Более того, в системе, обладающей развитым тезаурусом, появляется возможность автоматического получения логических выводов: Если некоторое явление сообщённое в документе характеризуется некоторыми дескрипторами, то это явление **должно** содержать характеристики, указанные как обобщающие дескрипторы и дескрипторы следствия, и **могут** содержать характеристики, указанные как видовые дескрипторы, ассоциативные дескрипторы причины. Если при этом на такую **возможность** указывают несколько дескрипторов, то эта возможность становится **высокой вероятностью.**

Возможность логического вывода свидетельствует, что тезаурус в системе представляет собой форму встроенного в систему **знания**, одной из разновидностей того, что сейчас называют **базами знаний**. А информационно-поисковая система, которая ищет документы, в которых не содержатся прями требуемые понятия, а они **следуют** логически из сообщаемых сведений, имеет полное право называться **интеллектуальной.**

Таким образом, на мой взгляд, естественный ход развития информационной теории и практики привёл нас к порогу, за которым встраивание в автоматизированные (компьютерные) системы «онтологий» в форме стандартных информационно-поисковых тезаурусов приведёт к технологиям **«искусственного» интеллекта**. И чем развитее будет встроенный в систему тезаурус, тем этот интеллект будет менее искусственным и более естественным.

5. Конечно структура ИПТ по ГОСТ 7.25 не является универсальной формой для баз данных. Но чтобы двигаться вперёд, целесообразно реализовать на практике то, что достигнуто теорией.

Так что же рекомендует нам ГОСТ 7.25, чтобы создать информационно-поисковый тезаурус? Во-первых, стандарт предполагает, что ИПТ создаётся для какой-то определённой ограниченной тематики, которую нужно определить в терминах общеизвестной классификации знания. Это в частности нужно для того, чтобы иметь возможность планомерно сочетать ИПТ разных предметных областей в суммарную картину мировой онтологии. Нынче имеется тенденция строить онтологии для всего языка, но эта задача благородная, но …

Далее следует набрать множество терминов, употребляющихся в данной области (словник). Это делается выявлением в документах, циркулирующих в нашей системе, ключевых слов – ограниченного числа терминов, более или менее описывающих тематику документа. И такая работа в настоящее время фактически делается во многих отраслях: Вы можете, листая библиотечную карточку, в низу библиографической карточки увидеть пару строк таких ключевых слов. Стандарт уточняет, какие именно КС могут входить в тезаурус; эти требования сводятся к тому, что они должны представлять вполне ясные понятия, могут быть однословными и словосочетаниями и даже компонентами сложных слов. Слова и словосочетания должны быть приведены к единому (словарному) виду – так, как их записывают в словарях. Многозначные слова должны быть снабжены пометами или комментариями, определяющими в каком смысле это слово употребляется здесь.

Примеры помет:

Стройка (процесс)

Стройка (место)

Проводник (вещество)

Проводник (человек)

Штамп (печать)

Штамп (инструмент)

Печать (пресса)

Печать (оттиск)

В качестве пояснения к термину может быть приписано его логическое определение.

На множестве собранных ключевых слов устанавливаются некоторые семантические отношения. Первое и необходимое отношение это – эквивалентность, **синонимия**. Ключевые слова, связанные отношением синонимии, считаются одним по смыслу элементом описания документов, который называется **дескриптор**. Расширительно этим термином называют любое КС, допущенное для вхождения в поисковые образы. Физически связь между ключевыми словами при представлении тезауруса на бумаге выполняется путём указания при данном слове всех его синонимов. Например:

Лингвистика

с: языковедение

языкознание

Из всей совокупности синонимов обычно только один фактически используется в поисковых образах в качестве представителя всей совокупности синонимов. Этот представитель также называется **дескриптором**. Так что этот термин имеет два значения, близких по смыслу. Те синонимы, которые не допущены до использования в поисковых образах, называются **аскрипторами**. В тезаурусе могут быть установлены также связи от аскрипторов к двум и более дескрипторам, заменяющим аскриптор либо альтернативно, либо совместно как словосочетание.

Кроме отношения синонимии в полноценном тезаурусе должны быть зафиксированы родо-видовые отношения понятий. Такое отношение устанавливается между двумя дескрипторами, если объём понятия одного входит а объём другого. Именно это отношение позволяет делать логические выводы абсолютно надёжно.

Стандартом также предусмотрено родственное отношение – не вхождение понятий, а вхождение друг в друга обозначаемых предметов. Это – отношение часть–целое. Например, слова «автомобиль» и «кузов» связаны этим отношением, но не отношением род–вид. Это отношение, важное, например, в производственной деятельности, не позволяет делать надёжные выводы. Так, из утверждения, что автомобиль движется со скоростью 100 км/час, следует это и для кузова. А вот из утверждения, что автомобиль стоит $ 10 000, не следует этого для его частей, которые дешевле.

Наконец, в стандарте предусмотрено установление связей между дескрипторами, значения которых «напоминают» друг друга. Это отношение **ассоциации**. В психологии различают два главных вида ассоциаций – по смежности и по сходству. И стандарт предусматривает возможность указывать какой вид ассоциации установлен в каждом конкретном случае. Ассоциация по смежности *(асм)* устанавливают между дескрипторами, когда обозначаемые ими предметы имеют **общие части** (например, общее

пространство). Ассоциация по сходству *(асх)* устанавливается, когда значения дескрипторов имеют **общие формы**.

Стандарт предполагает также возможность введения в тезаурус других отношений, важных для конкретной практики, при условии их точного описания.

С содержанием ГОСТ 7.25 необходимо познакомиться и знать предполагаемую им общую структуру информационно-поискового тезауруса.