1. 3 лекция. Информационно-поисковые системы и информационно-поисковые языки

0. На прошлой лекции мы обрисовали диапазон объектов, которые могут относиться к классу информационных систем и нарисовали схему системы научно-технической (деловой) информации. Теперь мы будем изучать её подсистему – **информационно-поисковую систему** – главную часть всей системы, которая поддаётся автоматизации, и автоматизация которой являлась стержнем развития социальной информатики вплоть до начала нынешнего века.

Замечание. Существует много (по крайней мере, три) конкурирующих представления о значении слова «информатика». Здесь мы будем оставаться в рамках изучения **социальной деловой** информации, как то изложено в Лекции 1 из курса введения в информатику, с которой я настоятельно предлагаю ознакомиться на моём сайте. Первоначально в нашей стране рассматривалась по преимуществу **научно-техническая** информация *(которая отвечала на вопросы **Что? Где? Когда?**)*, а теперь всё большее значение принимает **коммерческая** информация *(отвечающая ещё и на вопрос **Почём?**)*. Всё это вместе будем называть **деловой** информацией. И будем рассматривать системы обработки деловой информации. Сосредоточимся именно на информационно-поисковых системах. С проблемами, возникающими в связи с другими видами систем – управляющими – я предлагаю ознакомиться по реферату «Введение в конструирование АСУ» на моём же сайте. А связь проблем управляющих и информационно-поисковых систем подробно изложена в обзоре

> С. С. Терещенко. Проектирование автоматизированных систем научно-технической информации (аналитический обзор) // Итоги науки и техники. Серия «Информатика», № 4. – М.: ВИНИТИ, 1980. – 263 с.

1. Информационно-поисковая система **(ИПС)** имеет два входа. На один вход в неё поступают документы – либо в форме кратких библиографических записей, либо в форме полных текстов на естественном языке. (Здесь мы будем ограничиваться текстовым представлением данных; с другими формами вы знакомитесь, например в курсе геоинформатики). Документы систематизируются в форме каталогов. У другого входа ИПС стоит пользователь со своей **информационной потребностью**, сформировавшейся у него в голове. Чтобы осуществить поиск потребитель должен изложить свою потребность в виде объективизированного **запроса**. Обслуживающему персоналу системы (библиотекарю) можно изложить запрос на естественном языке; дальше он пойдёт просматривать каталоги по своему разумению. Но этого не достаточно для начала поиска в автоматизированной системе. Для автоматизированной системы необходимо представить запрос в форме, сопоставимой с принципами упорядочения документов в каталогах. Эту операцию библиотекарь делает в уме, а для автомата она должна быть выполнена в явной форме. Да и в случае живого библиотекаря вам, наверное, предложат заполнить формуляр, где указать, какая книга вам нужна. Этот формуляр и является объективизированным запросом, в принципе идентичным для автоматизированной и для ручной системы. В ручной системе формуляр рассматривает библиотекарь и достигает понимания Вашей потребности неформальным способом и сглаживает неточности и вольности ваших записей на формуляре. Так, если Вы неточно напишете фамилию автора (допустим *Гиляревский* вместо *Гиляровский*), то вам всё же выдадут книги о старой Москве, а не о теории информационного обслуживания.

Что же касается автоматизированного сопоставления вашего заказа с каталогом, то тут нужна большая осмотрительность. И не только в отношении орфографии. Нужны знания о принципах библиографического описания, поскольку каталог построен в соответствии с ними. Более того, нужны знания о многих детальных решениях, принятых именно в данном каталоге. Например, на какую фамилию стоят сведения о произведениях писателей, работавших частично под псевдонимом. Как искать, например, афоризмы Козьмы Пруткова? (Кстати, в алфавитом каталоге они обычно стоят на букву «К», а не «П»). Как искать произведения нескольких соавторов? Как искать труды нашего университета? на слово «МГУКИ», на «Московский государственный университет культуры и искусств», на «Библиотечный институт» (прежнее название, если нам нужны документы того периода), или как ещё? Всё это говорит о том, что надо знать **правила** составления запроса, учитывающие **правила** библиографического описания и составления каталога, особенно если каталог автоматический. Надо знать **слова, лексику**, которая для этого используется. Но если мы должны знать **слова и правила** их употребления, то значит мы должны знать некоторый **язык**. Этот язык называют **язык библиографических данных.** Правила этого языка закреплены стандартами на библиографическое описание (ГОСТ 7.1 – с ним вам необходимо познакомиться, он выложен на моём сайте). Они развиваются в методиках каталогизации литературы, изучаемых на библиотечных специальностях. Применительно к электронным каталогам и документам библиографические нормы истолковываются как стандарты на форматы машинных записей: ГОСТ 7.14 (ISO 2709), ГОСТ 7.19 – формат библиографических записей, ГОСТ 7.70 – описание электронных документов и баз данных. (С этими стандартами тоже нужно ознакомиться, хотя бы поверхностно. Если что будет непонятно, я разъясню на занятиях). Словарный состав этого языка складывается отчасти стихийно в виде имён авторов и заглавий работ, а отчасти регулируется специальными словарями нормативных написаний отдельных элементов библиографических данных (наименования стран, учреждений и др.), которые реализуются в библиотечных ИПС в виде так называемых «**авторитетных файлов**».

Не следует думать, что трудности поиска документов возникают из-за того, что *«злые библиографы нарочно путают пользователя, чтобы самим не остаться без работы»***.** Трудности поиска объективны, а библиографы и составители каталогов наоборот всеми силами стараются облегчить поиск, и именно для этого служат правила библиографии и каталогизации. В частности в каталогах библиографические записи по возможности дублируют, помещая их по алфавиту разных вариантов возможного поиска. Для традиционных карточных каталогов эта возможность ограничена физическим объёмом и трудозатратами на эту работу. При реализации каталога в виде машинных файлов в значительной мере проблема объёмов снимается, а трудозатраты на составление библиографических записей снижаются. Но чтобы преодолеть трудности библиографического поиска, нужно прежде всего знать что они существуют и знать их природу.

Несмотря на определённые трудности, поиск по библиографическим данным представляет собой довольно простую задачу, и реализуется в компьютерной поисковой системе достаточно просто. Для этого достаточно создать в компьютере так называемый «**инверсный файл**», или говоря по-иностранному, «**индекс**» (индексный файл). Этот файл состоит из упорядоченного перечня элементов библиографического описания, по которым может идти поиск (поисковые элементы), где каждому элементу сопоставлен адреса документов, имеющих этот поисковый элемент в своём библиографическом описании. Этот файл называется инверсным в отличие от списка первичных документов, в котором каждый документ характеризуется своим бибописанием, т. е. перечнем всех элементов бибописания. А в инверсном файле – наоборот, каждый поисковый элемент

сопровождается перечнем документов. Инверсный файл в компьютере упорядочивается так, чтобы программа поиска могла бы быстро обнаруживать в нём заданный элемент; конкретный способ упорядочения нам не существен, он задаётся программистами из соображений оптимизации программного обеспечения. Найдя в инверсном файле элемент бибописания, заданный пользователем, мы сразу получаем адреса релевантных документов и можем выдать пользователю их тексты. Инверсный файл является компьютерным аналогом библиотечной каталожной картотеки.

На инверсных файлах легко осуществлять поиск по сложным запросам, когда пользователь задаёт несколько условий поиска, связанных логическими отношениями. Например, *нужно найти произведения такого-то автора в соавторстве с таким-то соавтором и ещё произведения другого автора, изданные в таких-то издательствах*. Для исполнения такого запроса достаточно в список документов, полученных по одному поисковому элементу добавить адреса документов из другого списка и удалить из них адреса, отсутствующие в третьем списке. Такие операции называются **булевским поиском,** поскольку здесь моделируются операции **булевой алгебры (алгебры Буля)** – операции объединения, пересечения и другие комбинации множеств.

2. Более сложная задача стоит, когда пользователю нужно отыскать документы не по их внешним выходным данным, а по их содержанию. Впрочем содержание отчасти раскрывается заглавием документа, но заглавие редко становится поисковым элементом бибописания. Основным способом библиотечного раскрытия содержания является отнесение документа к той или иной области знания или сфере деятельности. Выбор областей знания основывается на философской классификации наук, а выбор сфер деятельности – на структуре общественной жизни. Но конкретный список классов, на который следует разделить документы, определяется практическими потребностями поиска того или иного содержания, а также объёмом документального фонда (чем больше фонд, тем дробнее должно быть деление). Так, для отыскания нужного учебника в школьной библиотеке достаточно их расставить по предметам школьного обучения (аналог классификации наук – математика, физика, химия, русский язык, литература, история, …) и по годам обучения (аналог структуры общественной жизни – 1-й класс, 2-й класс, ...).

В результате промышленной революции в конце XIX века возникло представление о научном знании как важной производственной силе и была поставлена задача инвентаризации всего накопленного человечеством знания. Для решения этой задачи бельгийские библиографы Поль Отле (Otlet) и Анри Лафонтен (Lafontaine) инициировали создание специальной международной организации, которая теперь называется Международная федерация по информации и документации (сокращённо – МФД, или ФИД от французского FID = Fédération Internationale d'Information et Documentation). В рамках этой организации была разработана и всеобщая схема классификации знаний, которая получила наименование «**Универсальная десятичная классификация**». С тех пор вот уже около 100 лет эта классификационная система успешно развивается и применяется для упорядочения (систематизации) библиотечных фондов и поиска в них литературы значительным числом библиотек и информационных служб.

Хотя УДК – не единственная система классификации знаний в информационных системах, мы остановимся не кратком описании её структуры, поскольку на этом примере можно уяснить все проблемы и приёмы организации информационно-поисковых языков **классификационного типа.**

Согласно УДК весь универсум знаний делится на 10 больших тематических полей (главных классов):

0 Общие вопросы науки и информационной деятельности

1 Философия, логика, психология

2 Религия, богословие

3 Общественно-экономические науки

4 (Свободный резервный класс)

5 Естественные и точные науки

6 Прикладные области знания (включая медицину, технику и сельское хозяйство)

7 Искусство, развлечения, спорт

8 Язык и литература

9 История и география.

Каждый класс в свою очередь делится на 10 (или менее) подклассов. Подклассы делятся дальше и дальше до любого необходимого уровня подробности. Обычны, например, классы девятого уровня деления, отражающие важные прикладные проблемы – квантовую электронику, защиту техники от коррозии и тому подобное.

Каждое деление обозначается десятичной цифрой, а цифры последовательных делений соединяются в одном индексе, где первая цифра обозначает номер деления на главные классы, вторая – номер подкласса первого уровня, третья – подкласс второго уровня, и так далее. Для облегчения зрительного восприятия индекса через каждые три цифры ставится точка.

Пример. Тема «Массы вещества, вынесенные на территорию обвалами и лавинами» имеет индекс УДК 551.435.644, где мы можем видеть следующую последовательность делений, постепенно уточняющих нашу тему:

5 – первая цифра индекса обозначает - естественные науки

55 – науки о Земле

551 – общая геология

551.4 – учение о формах земной поверхности

551.43 – отдельные формы рельефа

551.435 – формы рельефа, созданные внешними причинами

551.435.6 – формы рельефа, созданные силами гравитации

551.435.64 – аккумулятивные формы, созданные гравитацией

551.435.644 – формы, созданные падением обвалов и лавин.

Расшифровываются индексы УДК таблицами, полное издание которых занимает 10 томов средней величины (по 30 авторских листов, или по 200 страниц мелким шрифтом).

Кроме тематической характеристики УДК позволяет отразить в индексе некоторые дополнительные особенности документа или его содержания. Для этого в индекс добавляют определители этих особенностей, обозначенные специальными символами:

= – язык документа (=111 английский, =161.1 русский)

(=) – народ, к которому относится содержание документа: (=111) англоязычное население, (=161.1) русскоязычное население

(0) – форма, назначение документа (закон, учебник, справочник, или что-либо другое в этом духе)

(4/9) – страна, к которой относится содержание документа: (4) Европа, (470) Россия в целом, (5) Азия, (571) Сибирь и Дальний Восток России

« » - время, к которому относится содержание документа: «2005» нынешний год, «20» двадцать первый век, «19» двадцатый век

-0 – свойство основного предмета документа

.0

-1/9

‘1/9 – определители, значение которых раскрывается в таблицах применительно к каждому конкретному разделу.

Кроме того, допускается комбинировать разные классы для указания на документы, имеющие отношения к различным отраслям знания. Так что конкретный индекс УДК может иметь весьма сложную структуру. Например:

[55+622](470)(035)=/// – справочник по геологии и горному делу России

на английском языке,

где 55 - геология

622 – горное дело

(470) - Россия

(035) - справочники

=/// английский язык

Таким образом, характеристика документа индексом УДК читается и составляется действительно как языковое высказывание, в котором отдельные смысловые элементы (слова) при помощи вспомогательных знаков (препинания) по определённым правилам соединяются в единое целое, и число таких целых высказываний потенциально не ограничено.

Применение этого языка в информационной системе происходит неоднократно.

а) **На этапе создания** системы необходимо выбрать в УДК те классы, которые нам будут действительно интересны. При этом задача заключается не только в отсеве неинтересных классов, но также в конструировании комбинированных классов, которые точно выражают типичные информационные потребности, возникающие в работе нашей организации. Так, если мы работаем в системе Газпрома, то нас интересуют вопросы именно газообразных углеводородов, а в основных классах УДК этот аспект обычно не выделяется, поэтому мы должны включить в нашу рабочую классификацию сложные классы, в которых прочие характеристики соответствующих соединения будут дополнены признаком газообразного состояния.

б) **На этапе ввода документа** в систему документ требуется отнести к тому или иному классу каталога. Для этого нужно определить содержание документа, его тематику, и обозначить эту тематику теми или иными индексами УДК. Эта процедура называется ***индексированием документа***. Поскольку на основе полученного индекса будет строиться дальнейшее исследование документа, индекс должен по возможности отразить все темы, затронутые существенно в документе. Т.е. индексирование должно быть всесторонним, как того и требует международный стандарт ИСО 5963-85 и отечественный ГОСТ 7.66 и ГОСТ 7.59. Следовательно, и на этапе индексирования документов могут создаваться комбинированные классы, а документ может попасть в два и более классов заранее установленного каталога.

в) **На этапе индексирования запросов** выявляются те разделы каталога, в которых могут содержаться документы, необходимые пользователю. Эти разделы обозначаются соответствующими индексами УДК, которые также могут быть комбинированными, когда запрос не укладывается в заранее заготовленную сетку классов.

г) На этапе **сопоставления индекса** запроса с индексом раздела каталога может выявиться несовпадение индексов, которое, однако, не означает, что в хранилище нет необходимых документов. При комбинировании классов индексы могут на разных этапах вступать в комбинации в разном порядке и в разных сочетаниях. Следовательно, при сопоставлении индексов приходится их анализировать, выявлять в них элементарные составные части и их связи, т.е. нужно проводить что-то вроде разбора по членам предложения в грамматике. После такого разбора соответствующими запросу признаются документы, индексы УДК у которых не побуквенно совпадают с индексом запроса, а совпадают по своим значащим частям.

В автоматизированных системах эти процедуры, имеющие языковый характер, выполняются частично вручную (интеллектуально), а частично автоматически с помощью специально разрабатываемого программного обеспечения. Разработка программ конечно важное дело, но сперва должна быть осознана проблема и сформировано задание на программирование. Вот с этим подчас обстоит не всё благополучно. Зачастую программисты сами ставят себе задачу, не имея представления о проблеме в целом, и программируют не то, что истинно необходимо, а то, что они привыкли программировать.

3. В случае УДК, как видим, задача не так проста. Поэтому в 60х годах прошлого века, когда возможности компьютерных технологий были ограничены, разработчики первых автоматизированных ИПС пришли к выводу о нецелесообразности применения УДК. Требовалась более простая система классификации, без сложных правил, требующих интеллектуального подхода. И в нашей стране такая система была разработана и получила широкое распространение: Государственный рубрикатор научно-технической информации (ГРНТИ).

Схема ГРНТИ состоит из примерно 80 главных разделов, соответствующим научным дисциплинам и отраслям народного хозяйства. Они сгруппированы в 4 блока: Общественные науки; Точные и естественные науки; Технические науки (отрасли хозяйства); Комплексные проблемы. Каждый главный раздел поделён на подразделы, которых может быть до 100 (сантимальное деление). Обычно подразделов бывает от 5 до 20. Подразделы делятся на рубрики третьего уровня таким же образом. Дальнейшее деление рубрик не предусмотрено. Всего в ГРНТИ около 7000 рубрик, его издание занимает один том. При индексировании документов по ГРНТИ комбинирование рубрик не предусмотрено. Однако при необходимости один документ может быть отнесён к двум и более рубрикам.

4. Однако мы знаем, что 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 заголовком1[1] собираются библиографические описания документов. Составление предметных рубрик и распределение по ним документов называется **предметизацией**. Задача предметизации обычно состоит в том, чтобы указать главный предмет рассмотрения в документе и, может быть, основные его аспекты и 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, дополняющими друг друга, предназначенными для поиска документов по разным типам запросов. Предметизация даёт возможность собирать в одном месте документы по таким комплексам как материал, свойство, предмет (в узком смысле слова), явление природы или общества, род деятельности, географическое понятие и т. п., собирая под каждым предметным заголовком весь комплекс знаний, безотносительно к тому, какой области науки знания принадлежат.

Другим отличием предметизации от классификации служит то, что заранее составленный список предметных заголовков не ограничивает подробности анализа содержания документа. Если документ посвящён вопросу, не отражённому в списке предметных заголовков, имеется возможность сформулировать такой заголовок самостоятельно. Обычно же при выборе предметной рубрики для документа руководствуются заранее

1[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.

составленным списком предметных заголовков. 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. Пример фрагмента словаря предметных рубрик:

…

самовары

самолёты

- военные
- гражданские
- - грузовые
- - магистральные
- - пассажирские

см. также авиалайнеры

саморезы

см. винты-саморезы

…

Традиционная каталожная техника не позволяет раскрыть содержание документа предметными рубриками с приемлемой 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 предметных заголовков2[2]. Эти правила вместе со словарями рекомендуемых предметных заголовков составляют информационно-поисковые **языки предметных рубрик.**

Языки предметных рубрик используются не только как основа для предметного каталога, но также и как вспомогательное средство для пользователя систематическим каталогом, тематической классификацией. 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.

5. 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 всех документов, использовавших каждое ключевое слово.

2[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.

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 над записями инверсного файла, соответствующими ключевым словам запроса, должны производиться логические операции над множествами адресов документов аналогично тому, как это делается для сложных запросов на языке библиографических данных. Собственно новый способ описания и поиска документов – **язык ключевых слов** – сливается с языком библиографических данных в единый программный комплекс с едиными процедурами поиска.

Указанная организация языка ключевых слов позволяет сильно экономить на ресурсах памяти по сравнению с языком предметных рубрик. Так, если, допустим, документ заиндексирован пятью терминами, и мы хотим иметь возможность находить его по всем сочетаниям этих терминов, то в языке предметных рубрик мы должны были бы иметь не менее 80 предметных заголовков вместо 5 ключевых слов.

6. 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). 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.

Понятия могут быть заданы как совокупность всех способов обозначения их в текстах, 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:

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 включения в поисковые образы, а остальные синонимы используются как справочный материал 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 самый нормативный синоним, который в системе заменяет собой остальные синонимы. Последние получили наименование **«аскрипторы»** («не-описатели, не писомые»).

В дальнейшем было предложено вводить в 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, 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. Системы описания документов и запросов с помощью дескрипторов и информационно-поисковых тезаурусов называют информационно-поисковыми **языками дескрипторного типа, дескрипторными ИПЯ**.

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 «дескриптороспособность». 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 чтож! 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»?
